Smart Content Delivery in Cellular Network

Abbas Manavi Nezhad

School of Electrical Engineering

Thesis submitted for examination for the degree of Master of Science in Technology.

Espoo 05.10.2016

Thesis supervisor:

Prof. Olav Tirkkonen

Thesis advisor:

D.Sc. (Tech.) Kalle Ruttik

Aalto University
School of Electrical
Engineering

AALTO UNIVERSITY
SCHOOL OF ELECTRICAL ENGINEERING

ABSTRACT OF THE
MASTER'S THESIS

Author: Abbas Manavi Nezhad

Title: Smart Content Delivery in Cellular Network

Date: 05.10.2016 Language: English Number of pages: 7+50

Department of Communications and Networking

Professorship: Radio Communications Code: S-72

Supervisor: Prof. Olav Tirkkonen

Advisor: D.Sc. (Tech.) Kalle Ruttik

With the appearance of smartphones and a growing number of mobile subscribers, more media content is delivered to users. The growing volume of demanding media content causes challenges in usage of network resources. Deep understanding of changes in the temporal and spatial behavioral trends of users, can lead telecom companies and service provider to implement more optimized methods for resource allocation to the users.

One of the methods that can help is to schedule data delivery based on the user location and network statistics. In another words, predicting user mobility behavior and network resources can help to optimize data delivery scheduling. Anticipatory mobile communication is a field which uses different methods to provide more optimized decisions regarding content delivery, based on the prediction of the users' location and network resources.

In this thesis a system is modeled based on the user's location and available data rate prediction, by analyzing historical statistics of the network. Based on the assumption that a user's life pattern follows a cyclic daily pattern, following a Gaussian distribution, a Gaussian mixture model and expectation maximization (EM) algorithm have been used to predict the user mobility pattern in future. In addition, linear prediction is chosen to predict the network spectral efficiency and available bandwidth in future, through a regression matrix based on historical statistics. Applying this methodology, the available data rate to the user at different times and locations in future can be predicted. Merging this data with the user presence probability at different locations in future would result in deriving the probability distribution function of the required time duration to deliver a specific amount of data to the user.

Keywords: Anticipatory mobile network, Content delivery, Linear Prediction, Gaussian Mixture Model (GMM), EM algorithm

Acknowledgment

I would like to express my deepest appreciation to my supervisor, Professor Olav Tirkkonen, for providing me with the opportunity to conduct research on this subject, which opened a new horizon for me in my field of studies, and also for his kind support and patience during the course of this thesis.

A debt of gratitude is also owed to Dr. Kalle Ruttik, my thesis instructor, who has always been helpful to me by his guidance and invaluable comments.

I am very grateful to my family especially my mother and my aunt, Dr. Afsaneh Gächter, who have been a great source of inspiration to me throughout my education.

Finally, I want to present my appreciation to my real friends who have been always besides me throughout my studies in Finland, specifically during the past two years as the less pleasant years of my life.

I would like to extend my appreciation to Aalto University for providing me with the opportunity of doing my Master's degree studies.

Abbas Manavi Nezhad
Otaniemi, Espoo 05.10.2016

Contents

List of Figures

Abbreviations

3G	Third generation of mobile telecommunications technology
4G	Fourth generation of mobile telecommunications technology
bps	bits per second
BSS	Business Support Systems
CCDF	Complementary Cumulative Distribution Function
CDF	Cumulative Distribution Function
EDGE	Enhanced Data rates for GSM Evolution
EM	Expectation Maximization
FDD	Frequency Division Duplex
FIR	Finite Impulse Response
GB	Gigabyte
GMM	Gaussian Mixture Model
GPS	Global Positioning System
GSM	Global System for Mobile Communications
HD	High Definition
HLS	HTTP Live Stream
HTTP	Hypertext Transfer Protocol
Hz	Hertz
IIR	Infinite Impulse Response
IT	Information Technology
LBSN	Location Based Social Network
LTE	Long Term Evolution
OSS	Operations Support System
PDF	Probability Density Function
PMF	Probability Mass Function
QoS	Quality of Service
SNR	Signal to Noise Ratio
TDD	Time Division Duplex
UHD	Ultra High Definition

1 Introduction

1.1 Motivation

Running business in today's growing industrial world, requires creative and more optimized solutions. One of these solutions can be acquisition and analysis of real-time data about changes in customers' demands. Telecommunication Industry is one of the largest types in service provision section of economy. The influence of the IT and telecom services on daily life of people all around the world, has led to a saturation of this industry by by small and big service providers during the past decades. Concerning the current trend of the market, providing services alone is no longer profitable and the key to success is value added innovative solutions. There are significant fluctuations in customers' behavior timeline of using the services. Collecting users' behavioral data and matching them with existing resources can result in understanding traffic demand, which has a significant effect on effective marketing and service provision solutions, and subsequently enhance users satisfaction rate by higher QoS and more affordable prices [11].

One of the biggest challenges in telecommunication is the optimized usage of bandwidth and resources. Different methods and algorithms have been created for this purpose but utilizing the bandwidth and resources even in the best optimized system has a limit. Overloaded networks at peak times face call dropping and decreasing of the QoS on the running services. But how can we deal with overloaded networks?

One of the solutions is to motivate users to transfer a portion of their data usage from traffic peak time to the times on which the network is not fully used. According to how busy the network might be at a time, the excess resources can be used to motivate users for using the network at that time, instead of the rush hours. This can be done by offering discounted data packages or extra speed, at less loaded time of the network.

Deep understanding of changes in users temporal and spatial behavioral trends, can lead telecom companies and service provider to a profitable pricing policy [11].

On the other hand, progress and advancement in cellphone hardware technology, have made it possible for the users to store and watch high quality images and videos, which is the complement of higher wireless connection data rate provided by 3G and 4G networks. Thanks to these technologies, smart phone users' demand for big volume data such as HD movies is rising [12].

By looking at the present and near future of mobile users subscribers and data traffic statistics, the urgent vital demand for more optimized bandwidth usage solutions is clear. For example the monthly mobile traffics predicted to undergo a 727 % growth in 2020 comparing to 2015. Figure 1 shows the Cisco estimation of monthly

raise in mobile data traffic in Exabyte from 2015 to 2020 [1].

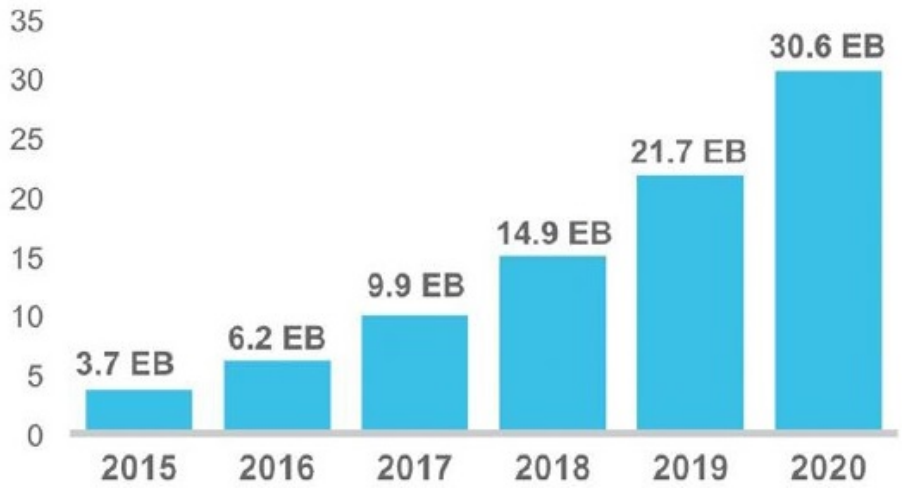

Figure 1: Cisco forecast for monthly usage of mobile data traffic by 2020 [1].

Moreover, the forecast of used technology, shows a significant shift from older radio access technologies such as GSM/EDGE to new advanced technologies such as 4G or LTE. Figure 2 represents Ericsson mobile subscription estimation by technologies from 2011 to 2021 [2].

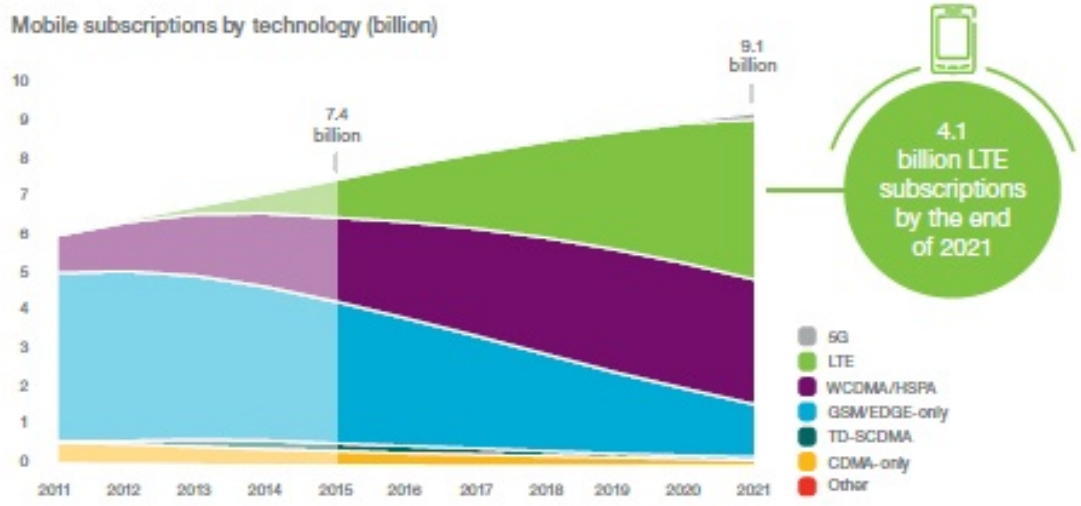

Figure 2: Ericsson estimation of mobile subscriptions by technologies [2].

1.2 Objective of Thesis

The goal of this thesis is to design and simulate a model for estimating the probability of total transferable bits to the user in the future. This is done by predicting the user's location and network available resources in the future based on the historical data. This probability can be used to find the estimated time which is needed to transfer a particular content, such as a video clip, with a specific volume.

This method and its output can be considered in network tariff policies and marketing purposes or in OSS/BSS systems by improving real-time decision making [13].

1.3 Structure of Thesis

The second chapter of this thesis is a short diary about the idea and solutions of anticipatory mobile networks. In the third chapter, a parametric model is derived. This model is used to create the simulation model in Chapter 4 based on the defined scenario. In Chapter 5 the experimental results of the simulated model are presented to demonstrate the prediction abilities of the system.. The last chapter is a summarized conclusion and suggestion for continuing the presented method in this thesis.

2 Anticipatory Mobile Networks

During the past decade, lifestyle of millions of people have been revolutionized by mobile phones. In 2015, 3G mobile network was able to cover 69% of the world's population [14]. By appearance of smart phones, another significant change happened in people's lifestyle. Nowadays, the presence of smartphone including their topnotch hardware technologies along with people with almost every step they take, make them a pocket-sized powerful device which can do many tasks and computations like a desktop computer.

Different types of sensors are implemented in smartphones, they can measure and sense movement, location, moisture, even heartbeat and biometric characteristics. These sensors accompanied by other processing capabilities of smartphones, make them statistic recording devices which can sense and record the historical social and temporal behavior of the users. It is possible to assume the mobile communication network aa s large scale human behavior sensor [15].

Anticipatory mobile communications is a field that proves the essentiality of mobile sensing and machine learning for more optimized and smart decision making solutions, based on the prediction of the future statistics. For more accurate and efficient prediction, a solid relation between the user, the device and the environment should be created. Smartphones are a suitable platform for anticipatory mobile computing as they can base their decisions and actions on a predicted model of their state and the environment [3].

Different required processing stages in anticipatory mobile computing and the relationship among them is illustrated in Figure 3. In this figure it is shown that, the context is sensed, modeled and predicted by the mobile phone and feedback from the user, specifies if the anticipatory decision are accomplished. All the processes are distributed between the cloud and the mobile phone [3].

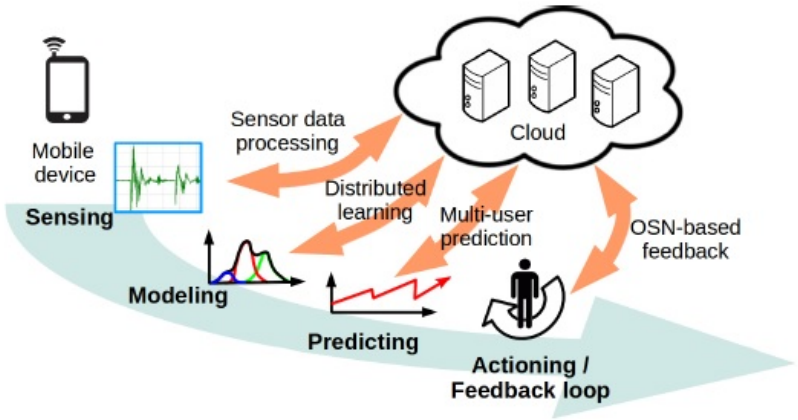

Figure 3: Anticipatory mobile computing architecture [3]

Anticipation of the user's behavior and network resource state, can be a key for more efficient and optimized resource allocation and utilization of the network. Anticipatory mobile network can be implemented by understanding and modeling of the user behavior. Behavioral model is a key approach for anticipation [16].

Anticipation can be an effective approach for efficient dynamic resource allocation. Measuring and analyzing network statistics can give a clear image of the dynamic resource allocation essentiality. As an example, in [17] the mobile subscriber behavior based on their generated traffic, mobility and activity have been studied on a temporal scale. The interesting key observations of that study are user's traffic load and mobility. This study shows 90% of the generated traffic are for less than 10% of users, while more than 50% of the load is handled by 10% of base stations. In terms of mobility, majority of the users have cyclic temporal behavior and it is highly probable that they be present at the same time at the same location every day.

The main goal in radio resource management is the most efficient utilization of the radio resources to guarantee the demanded Quality of Service (QoS) for each user by allocating the sufficient resources to each user. Moving users are a big challenge in radio resource allocation, specially the users whose radius of movement includes different cells, such as users in vehicles. User movement cause fluctuation in the experienced bitrate of the user via data connection [18].

Many people during their daily trips use public transport and majority of them entertain themselves with their smartphones through the trip and streaming videos are among the most popular content used by smartphone users. Assume a user in a bus watching a video on a smartphone. As the bus direction changes through the trip, user experiences different bitrates and poor playback quality in area with lower data rate coverage. Many solutions have been used to keep the video playback on smartphones smooth, such as scalable video codec or segmented adaptive HTTP streaming [6]. In adaptive HTTP streaming, the playback time is divided into HTTP-based small interval segments, encoded at different qualities, and the client player automatically downloads the segments based on the condition of the network. ??? If the user path can be predicted, network can allocate more resources to pre download the higher quality video segments before the user goes to a low rate area. It can be a solution for buffered streams. On the other hand, network is also able to allocate more resources to users who just enter a high rate area from a low rate area, to speed up the buffering for corrupted playbacks. More accurate bitrate prediction can results in better and higher video playback [18].

As it was mentioned before and also it is shown in [19], users daily routes have a high level of temporal and spatial correlation which causes a correlation between a specific location and received signal strength at that location. It can be used to create radio maps. Figure 4, shows and example radio map of Helsinki and Espoo for Elisa mobile operator.

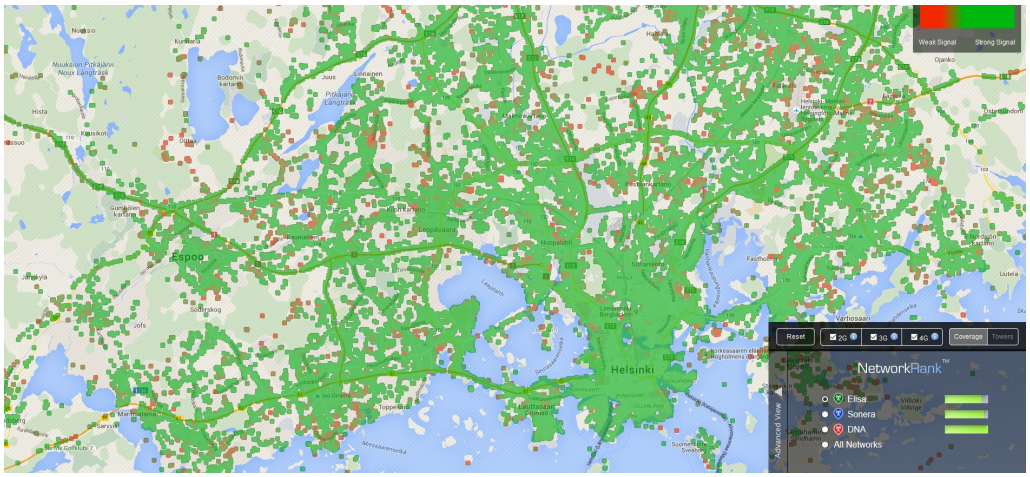

Figure 4: Helsinki and Espoo radio signal strength map for Elisa network [4]

In [5], wireless link quality prediction based on the geographical path loss map has been studied. The anticipation base in that study is the intersection nodes in the streets, which are the points on the map where a user may changes the direction. Figure 5 shows an example of anticipation scheme for user trajectory prediction [5].

The algorithm in Figure 5 is the combination of two phases. First phase is the offline phase which is finding and analyzing intersection nodes from the street map data. Second phase is real time prediction phase which refers to real time prediction for the estimating arrival time of the user to the next intersection point, when its current location and direction is known. The phase 1 information about street maps are the input for the phase 2.

Anticipatory algorithms can be used to predict the variation in available resources to the user through the user movement and it can be used for scheduling to transfer more data to the user when the bit rate is high with the aim of pre-downloading HTTP segments to buffer. This can prevent delays or rebuffing in downloading segments when user is crossing locations with lower data rate coverage and prevent playback quality decrease [20].

Experienced bitrate by a smartphone data connection is impressed by geographical location as not all locations have the same radio resource and signal coverage; therefore, user direction prediction along with radio maps coverage is one of the most important factors which can be used to deliver demanded content to the user in a more efficient way. One of the solutions for user direction prediction is based on GPS data which can be used by the network to schedule video segment transfer to the user based on a GPS-based bandwidth table. Bandwidth monitoring helps the network to lookup the available bitrate through predicted direction of the user, which is based on the GPS data, and make decision about the amount of data segments

Algorithm 1 Anticipation Algorithm

Require:
ListOfStreets
ListOfIntersectionNodes

```
1: while User is moving do
2:     UserPos = GetUserPos()
3:     Direction = GetUserDirection()
4:     UserStreet = GetUserStreet()
5:     INode = PredictIntersectionNode(UserStreet,Direction)
6:     UserSpeed = PredictUserSpeed()
7:     Distance = DistanceFromCurrentPositionToINode
8:     ETA = DistanceToIntersectionNode / UserSpeed
9:     PredictedUserPos = PredictUserPositions(ETA)
10:    PredictedPL = GetPLFromCM(PredictedUserPos)
11:    Return PredictedPL
12:    Wait until user crosses the intersection node
13: end while
```

Figure 5: Anticipation algorithm for route prediction [5]

to be buffered to the user during different times in the trip. Figure 6 illustrates a GPS-based bandwidth lookup architecture [6].

Figure 6: GPS-based bandwidth-lookup system architect [6]

(1) The receiver selects a suitable route to the destination and sends it to the bandwidth database. Additionally, the receiver records the routes and path prediction is based on receiver's travel history.
(2) When the bandwidth is received by the bandwidth database, it sends a sample of historic recorded bandwidth through the receiver's chosen path. This sample includes the average bandwidth and standard deviation.
(3) The receiver combines the bandwidth samples from bandwidth database and the estimated travel time based on historic recorded data. By this information, receiver can calculate the estimated amount of downloadable data at different geographical locations during the trip. This information helps the receiver to plan the data segments download scheduling during the trip.

There are studies about different anticipation buffer control and resource allocation such as [7], [21], and [8] for scheduling efficient delivery of video contents. These works are all based on HTTP Live Stream protocol (HLS), in which video is transferred in form of divided video segments instead of a continuous stream. One of the options in HLS is that each video segment can be stored on the server with different quality and size and based on the link quality of the user, the network can decide which video segment with which quality, is the best option to be delivered to the user. Existing approaches to video segments delivery through HLS protocol based on the anticipated data rate can be divided into long anticipation horizon and short anticipation horizon [8].

In long anticipation horizon, it is assumed that the data rate is anticipated for the complete duration of the video. One of the algorithms that can be used in long anticipation horizon is Fill Algorithm, which is designed to prevent fixed buffer size shortcomings [7]. This algorithm uses each time slot's data rate for input parameters, repeating over all the time slots and in case of availability of required data rate, buffer the video segments to the user with the best quality. Otherwise the algorithm decreases the video quality by buffering more video segments with lower quality.

Figure 7, shows the main structure of Fill Algorithm. ANTICIPATEUSERRATES(u) returns a user's anticipated data rates for all time slots considering channel anticipation and radio resource scheduler. Figure 8 represents the SCHEDULESEGMENT. Based on the available data rate in the time slot, there are two possible different operations.

Algorithm 1 FILLSCHEDULER(U, T, Q)

```
1: // users U, times T, qualities Q
2: for all u ∈ U do // schedule all users
3:   C ← ANTICIPATEUSERRATES(u) // from channel anticipation
4:   s ← 0 // initialize couter for scheduled segments
5:   for all t ∈ [0..⌈T⌉] do // schedule all timeslots/segments
6:     s ← s + SCHEDULESEGMENT(u, t, s, Q, C)
7:   end for
8: end for
```

Figure 7: Fill Scheduler Algorithm [7]

If there is sufficient data rate at the current time stamp for downloading a new video segment, the Fill algorithm downloads the maximum quality available video segment. This operation keeps the minimum number of video segments in the buffer as long as there is no need to pre download video segments for future time slots with

low data rate. If the Fill algorithm, during the iteration process, finds the anticipated data rate in time slot t insufficient for downloading even the lowest quality video segment, the schedule for one or more time slots before time slot t has to be changed to pre download more video segments before time slot t [7].

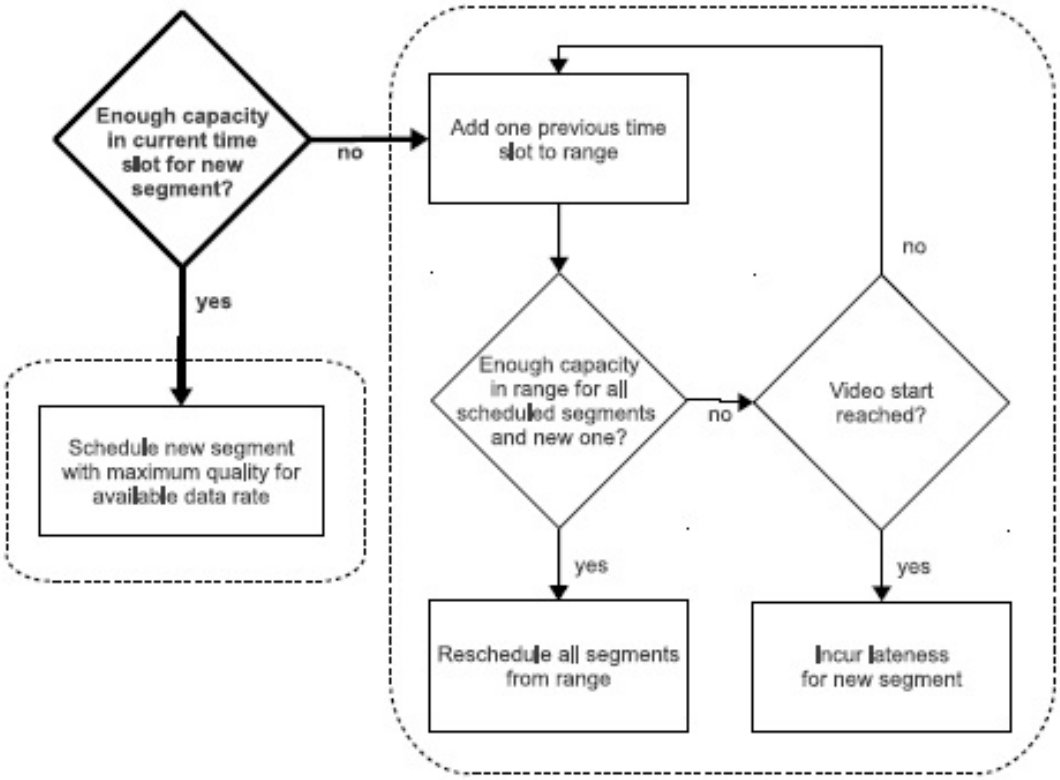

Figure 8: Fill Scheduler flowchart [7]

The following helper functions are used in the algorithm 2 in Figure 9.

- getBestQuality(Q, c)
  Based on the anticipated available data rate c, returns the best available quality of the video segment (out of Q), and for insufficient data rate even for lowest quality video segment, returns FALSE

- getBestQualityRange(Q, n, c)
  Based on the available data rate c and the n segments to be downloaded, returns the best possible quality of video segment (out of Q), and for insufficient data rate even for lowest quality video segment, returns FALSE

Algorithm 2 SCHEDULESEGMENT(u, t, s, Q, C)

```
1: q ← GETBESTQUALITY(Q, C[t])
2: if q ≠ false then // enough capacity in current time slot for
   new segment?
3:  SCHEDULE(u, s, t, q) // schedule new segment with maximum
    quality for available data rate
4:  return 1
5: else // even lowest quality not feasible in time slot t
6:  for all g ∈ [t..0] do
7:   // enough capacity in range [g..t] for all scheduled segments
     and new one?
8:   if GETBESTQUALITYRANGE(Q, t − g + 1, Σ_{i=g}^{t} C[i]) ≠ false
     then // going back to g provides enough data rate
9:    q ← GETBESTQUALITYRANGE(Q, t − g + 1, C[g...t])
10:   p ← 0
11:   for all r ∈ [g..t] do // reschedule all segments from range
12:    n ← GETSEGMENTSFORQUALITY(q, C[r])
13:    for all v ∈ [(g + p)..(g + p + n)] do
14:     SCHEDULE(u, v, r, q)
15:    end for
16:    p ← p + n
17:   end for
18:   return 1
19:  end if
20: end for // video start reached
21: return 0 // incur lateness for new segment
22: end if
```

Figure 9: ScheduleSegment Algorithm [7]

- GETSEGMENTSFORQUALITY(q, c)
  Considering the available data rate c and the quality q for video segments, returns the number of downloadable segments

- SCHEDULE(u, s, t, q) Schedules the download of segment s for user u at time t with quality q

In short anticipation horizon, the assumption is that predicted data rate is only available for a limited number of future time slots. Figure 10 represents the short horizon algorithm. In this algorithm, c_c is the data rate for the current time slot and C_p is the anticipated data rate for the future time slots. The mechanism for this algorithm is that it downloads additional video segments at the current time slot if it will not be possible to download video segments in future time slots. s and k are the number of downloaded video segments and number of played segments so far, respectively [8].

Algorithm 3 Anticipation Algorithm

```
Require:
  ListOfStreets
  ListOfIntersectionNodes
1: while User is moving do
2:     UserPos = GetUserPos()
3:     Direction = GetUserDirection()
4:     UserStreet = GetUserStreet()
5:     INode = PredictIntersectionNode(UserStreet,Direction)
6:     UserSpeed = PredictUserSpeed()
7:     Distance = DistanceFromCurrentPositionToINode
8:     ETA = DistanceToIntersectionNode / UserSpeed
9:     PredictedUserPos = PredictUserPositions(ETA)
10:    PredictedPL = GetPLFromCM(PredictedUserPos)
11:    Return PredictedPL
12:    Wait until user crosses the intersection node
13: end while
```

Figure 10: Short anticipation horizon algorithm [8]

Below is the description for the helper functions used in the algorithm.

- BESTQUALITY(c)
  Returns the best downloadable quality for a segment with anticipated available data rates c, or -1 if there is not enough data rate even for the lowest quality

- BESTQUALITYMULTI(c, n)
  Returns the best possible quality in which n segments can be downloaded with anticipated available data rate c

- SCHEDULE(s, q)
  Schedules the download of segment s with quality q for the current time slot.

These algorithms are part of the literature review about anticipatory mobile networks and in this thesis no algorithm will be used.

3 System Model

Many studies are performed to model users' mobility behavior with different aims such as improving location-based services [22][23] and more efficient mobile marketing [24]. The goal of this thesis is to calculate the probability of the estimated time to deliver a specific amount of data to the user in future based on the user's historical behavioral data and network statistics. To reach this goal, user location at each time slot and available bitrate to the user at that location should be anticipated. System model which is used in this thesis is based on two prediction parts. First part is the prediction of the user's location on the network at each time slot in the future, i.e. users mobility behavior. The second part of the system model is prediction of the total available bits to the user at each time slot in the future. Free radio resource and spectral efficiency are the network parameters which are determinative of the deliverable bits to the user at each time. Therefore, the second part of the system model consists of predicting the two aforementioned parameters at each location at each time.

In this chapter, first a framework for modeling user mobility behaviour to predict the user location in future, is defined. Then, a model to predict total available bits to the user at each time and location in the future will be derived, and after that a framework for calculating the probability of the total available bits to the user is presented. After these stages, an approach to find the probability of the time required to deliver a specific amount of data to the user in the future will be defined.

3.1 User Location Prediction

At first, to design a model, the required parameters should be described. For determining the parameters, it should be known what specifications are required to be modeled. For modeling the temporal pattern of the user, recorded historical data of user presence at each location and time stamp, are key parameters. Let us assume $U = \{u_1, u_2, u_3, ..., u_n\}$ is the set of users and n is the number of users. $L = \{l_1, l_2, l_3, ..., l_m\}$ is the set of locations and m is the number of locations. $T = \{t_1, t_2, t_3, ..., t_k\}$ is the set of time stamps, and each member of this set, indicates a specific time in historical data. Based on the defined sets, each observation is described by three parameters, user, location and time, which can be represented as $< u_i, l_j, t_z > \in O$. O is the set of observations in historical data. $H_{u,t}$ is the set of observations which belongs to user u_i at location l_j before time t.

Given observed historical data $H_{u,t}$, the presence probability distribution of the user u_i, at location l_j at time t_z can be presented by [25]:

$$P(c_u = l_j \mid t_z, H_{u,t}) \tag{1}$$

where

$$H_{u,t} = \{< u_i, l_j, t_z > | < u_i, l_j, t_z > \in O, u_i = u, t_z < t\}. \tag{2}$$

and c_u is the presence location of the user.

Human mobility has a cyclic temporal continuity characteristic. For example a user usually spend the working hours (8:00 -16:00) of the day around a specific geographical area. This characteristic makes it possible to guess unobserved temporal specification using historical temporal patterns, which are the observed human mobility pattern [25] [22] [23]. This cyclic temporal patterns can be used for user temporal pattern prediction.

Since user mobility behavior can be considered as an stochastic process[25], temporal features of user's mobility pattern are sparse in the large temporal feature space. Therefore, prediction accuracy can be prone to an impressive effect due to unobserved features. As user's mobility pattern has a temporal continuity, unobserved temporal features can be estimated from observed temporal pattern [25].

For example, based on the observed temporal data, user had been at a specific library at 17:00 and 18:00 many times, but no data recorded for user presence at 17:30 at that library. This pattern can indicate a high probability of the user's presence at that library at 17:30 in the future, even if the user had not been observed at 17:30 in that library in historical observation data [25].

Applying Bayes' rule, the Eq.(1) can be reformed to

$$P(c_u = l_j \mid t_z, H_{u,t})$$
$$\propto P(t_z \mid c_u = l_j, H_{u,t}) P(c_u = l_j \mid H_{u,t}) \tag{3}$$

$P(c_u = l_j \mid H_{u,t})$ is the location distribution or spatial state of the user u_i by knowing his/her historical presence data without considering temporal information. $P(t_z \mid c_u = l_j, H_{u,t})$ is the temporal state distribution of the user u_i, knowing his/her observed presence data at location l_j [25] [26].

Before predicting the user presence probability, it should be known how it is possible to find a solution to model temporal data of the user at each location. By considering temporal state of the user in one specific location and the assumption that the probability of the user presence at time t_z is only relevant to the location of the user at time t_z [26]:

$$P(t_z \mid c_u = l_j, H_{u,t}) = P(t_z \mid c_u = l) \tag{4}$$

Following the above consideration, Eq.(4) can be rewritten as

$$P(c_u = l_j \mid t_z, H_{u,t})$$
$$\propto P(t_z \mid c_u = l_j, H_{u,t}) P(c_u = l_j \mid H_{u,t})$$

$$= P(t_z \mid c_u = l)P(c_u = l_j \mid H_{u,t}) \tag{5}$$

Figure 11 is a sample of visiting frequency of the user at various times of the day at a specific location and Figure 12 is the corresponding estimated Gaussian distribution of user presence probability at location l_j. As it can be seen, the figure has a distribution very similar to Gaussian distribution. This conclusion is the result of the analysis in [26] which is done on users data provided by Nokia Mobile Data Challenge.

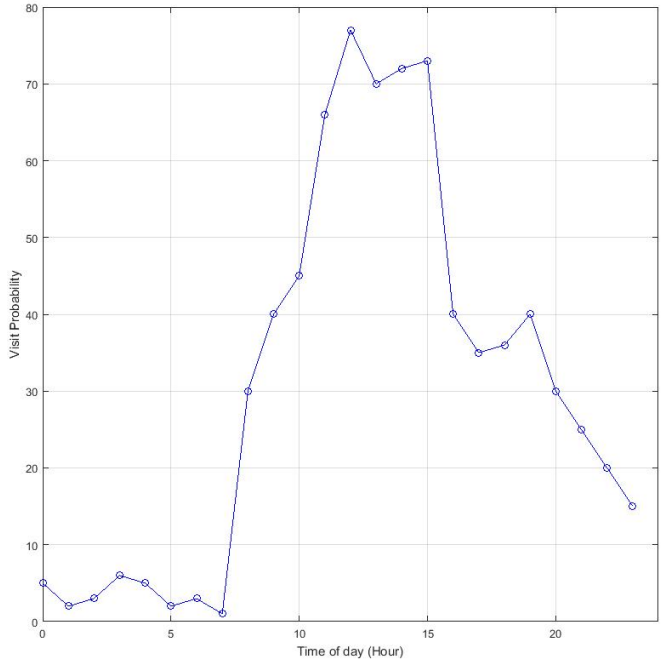

Figure 11: User visit frequency at different times at location l_j during a day

By this knowledge, it can be assumed that user presence at a specific location during a day can be modeled by a Gaussian distribution. Hence, to find the presence probability of the user at a specific time and a given location a Gaussian distribution can be used [26]:

$$P(t_z \mid c_u = l_j) = \mathcal{N}_L(t_z \mid \mu_{t_z}, \sigma^2_{t_z}) \tag{6}$$

where μ_{t_z} and σ_{t_z} are the mean and variance of the Gaussian distribution. By knowing the μ_{t_z} and σ_{t_z}, the joint probability of historical data can be written as [26]:

$$P(t(h) \mid \mu_{t_z}, \sigma^2_{t_z}, l) = \prod_{z=1}^{V_L} \mathcal{N}_L(t_z \mid \mu_{t_z}, \sigma^2_{t_z}) \tag{7}$$

$$= \prod_{z=1}^{V_L} \frac{1}{\sqrt{(2\pi\sigma^2_{t_z})^{\frac{1}{2}}}} exp\{-\frac{1}{2\sigma^2_{t_z}}(t(h) - \mu_{t_z})^2\} \tag{8}$$

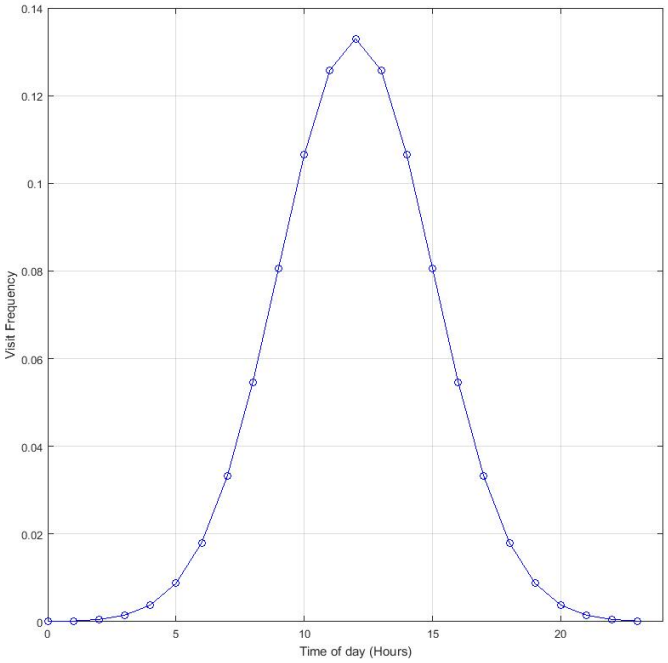

Figure 12: The smoothed user visit probability at different times at location l_j during a day

V_L is the total number of the visits conducted by the user at location l_j. t_z is the time in which the visit has happened. For obtaining the μ_{t_z} and $\sigma^2_{t_z}$, the above log likelihood should be maximized. For this purpose, the EM algorithm can be a suitable solution [27].

3.1.1 Gaussian Mixture Model and EM Algorithm in Location Prediction

A Gaussian Mixture Model (GMM) defines a probability density function which is created by summation of weighted Gaussian components. This powerful model has attained great popularity in many engineering and science disciplines, such as speech recognition and soft clustering. Besides, GMM is an effective solution to estimate a function from data where two or more Gaussian density components are mixed in different proportions. This property of Gaussian Mixture Model gives it the capability to appropriately describe many types of physical data, and thus making it possible to decompose the Gaussian mixture distribution into context dependent components.

Suppose random variable $\mathbf{x}$ has a Gaussian mixture distribution. Its PDF is determined by weighted superposition of Gaussians as

$$p(x) = \sum_{k=1}^{K} P_k \mathcal{N}(x \mid \mu_k, \sigma_k) \tag{9}$$

where P_k represents mixture weight of k^{th} Gaussian component that must satisfy the following condition and sum up to unity,

$$\sum_{k=1}^{K} P_k = 1$$

and $\mathcal{N}(x \mid \mu_k, \sigma_k)$ denotes Gaussian distribution parametrized by (μ_k, σ_k). Moreover, the number of mixtures is determined with accordance to the structure of the data and problem. For example, AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion)are two techniques developed based on information from theoretical approach to determine the number of mixtures in GMM. However, it is possible to use customized techniques based on the problem to be solved and the type of data to be modeled. The technique used in this thesis will be explained later.

By formulating the Gaussian Mixture Model in terms of discrete random variables, it is feasible to solve GMM problems via EM algorithm, where the model depends on unobserved latent variables. Assume a binary random variable Z_K, the latent variable is embedded in the GMM model, where $Z_k = 1$ and the other components are equal to zero. Clearly z_k elements are drawn from $0, 1$. It can be inferred that mixing weights are related to random variables z as shown in the following equation:

$$P(z_k = 1) = P_k$$

where

$$sum_{k=1}^{K} z_k = 1$$

Also the probability distribution of $P(z)$ is written as:

$$P(z) = \prod_{k=1}^{K} P_k^{z_k}$$

and the condition distribution of x, given z_k is a Gaussian written as:

$$P(X \mid z) = \prod_{k=1}^{K} \mathcal{N}(x \mid \mu_k, \sigma_k)^{z_k}$$

As a result, the probability distribution of x can be obtained by marginalization of $P(x, z)$ in the form of

$$P(x) = \sum_{Z} P(z) P(x \mid z) = \sum_{k=1}^{K} P_k \mathcal{N}(x \mid \mu_k, \sigma_k)$$

In order to estimate the parameters of Gaussian Mixtures, the log likelihood is maximized with respect to parameters including means, covariances and weighting coefficients. One powerful and elegant method to find these parameters of Gaussian mixture model is Expectation Maximization (EM) algorithm. EM algorithm can be used to find the maximum likelihood for models with latent variables. Dempster [28], formulated the convergence and proposed the term Expectation Maximization (EM) algorithm. EM algorithm has been widely utilized in many areas such as signal processing, pattern recognition, and econometric where latent variables affect outcomes of physical phenomena.

A major complication in majority of parameters estimation problems is the impossibility of direct access to the data or that part of the data which is missing. The EM algorithm is ideally developed to deal with the problems of this kind. The EM algorithm estimates parameters in an iterative approach with two steps:

1. **Expectation step**: The expected value of the log likelihood function is computed with respect to z given x, through the current estimate of unknown parameters θ and observed data.

$$Q(\theta \mid \theta^{(t)}) = E_{z|x,\theta^{(t)}}[\log L(\theta; x, z)] \tag{10}$$

2. **Maximization step**: the maximum likelihood estimate of the parameters is computed using the data provided by the expectation step.

$$\theta^{(t+1)} = \arg\max Q(\theta \mid \theta^{(t)}) \tag{11}$$

The strong capacity of Expectation Maximization algorithm in dealing with estimation problems with latent variables, brings this idea that it can resolve the estimation problem of Gaussian Mixture Models. In addition, it must be emphasized that EM algorithm does not guarantee the convergence to global maximum of log likelihood when there are multi local maxima.

The necessary steps to perform EM algorithm on Gaussian Mixtures are:

1. Initialize the means μ_k, covariances σ_k and weighting coefficients p_k and evaluate the initial value of the log likelihood.

2. **E Step**: Compute conditional probabilities $P(z_k = 1 \mid x_n)$ using the current parameter estimates

$$P(Z_k = 1 \mid x_n) = P_k \frac{\mathcal{N}(x_n \mid \mu_k, \sigma_k)}{\sum_{j=1}^{K} P_j \mathcal{N}(x_n \mid \mu_j, \sigma_j)} \tag{12}$$

3. **M Step**: Compute the new parameter estimates using the current conditional probabilities

$$\mu_k^{new} = \frac{1}{N} \sum_{n=1}^{N} P(Z_k = 1 \mid x_n) x_n \tag{13}$$

$$\sigma_k^{new} = \frac{1}{N_k} \sum_{n=1}^{N} P(Z_k = 1 \mid x_n)(x_n - \mu_k^{new})(x_n - \mu_k^{new})^T \tag{14}$$

where

$$N_k = \sum_{n=1}^{N} P(Z_k = 1 \mid x_n)$$

4. Evaluate the log likelihood

$$Eval(\mu^{new}, \sigma^{new}, P^{new}) = \ln P(X \mid \mu^{new}, \sigma^{new}, P^{new}) =$$

$$\sum_{n=1}^{N} \ln\{\sum_{k=1}^{K} P_k^{new} \mathcal{N}(x_n \mid \mu k^{new}, \sigma k^{new})\} \tag{15}$$

and check whether the convergence condition is satisfied, if not return to step 2.

User's mobility pattern during a day shows a periodic movement between specific most visited places, such as home and work, during certain time durations [26]. Comparing this definition with the assumption of the user mobility pattern at a location during a day, which is similar to a Gaussian distribution and is illustrated in Figure 12, it can be more accurate if the user mobility pattern modeled by mix of Gaussian distributions during a day. Each one of these Gaussian distributions is centered at a certain time duration. For more accuracy in prediction of the user presence probability at a specific location and time in future, user temporal pattern in a day can be assumed as a mixture of K Gaussian distributions, with each one being simulated by Eq.(9) with its own specific μ and σ [25] [27].

After finding μ and σ for each of K Gaussian distributions, it is possible to regenerate the user presence probability during one day at one specific location l using Eq.(17)[25].

$$P(t_z, c_u = l_j, H_{u,t}) \sim \sum_{i=1}^{K} A_i \mathcal{N}(t_z \mid \mu_{i,l_j}, \sigma_{i,l_j}) \tag{16}$$

K is the number of Gaussian distributions, whose mixture generate the user temporal pattern in one day. A_i is the i^{th} Gaussian distribution maximum power. μ_{i,l_j} and σ_{i,l_j} are the mean and variance of the i^{th} Gaussian distribution at location l_j, respectively.

Figure 13 presents another example of smoothed Gaussian distribution of user presence probability at a different time during a day and at one specific location l_j. From the figure, it can be assumed that the user presence probability can be

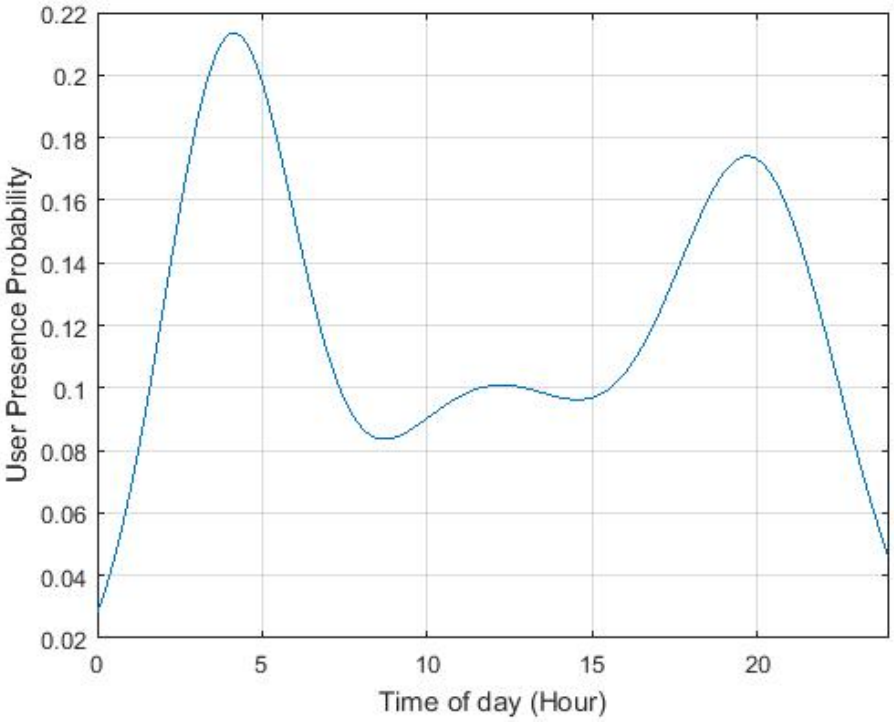

Figure 13: The smoothed user visit probability at different times during a day

approximated with mix of three Gaussian distribution [25].

Figure 14 shows an example of three Gaussian distribution that can be Gaussian distributions compositions of Figure 13, each one with its own specific μ and σ. In this example $K = 3$.

By extracting μ and σ for each of these Gaussian distribution components, user presence probability at each location can be modeled. This model includes the user cyclic movement pattern.

To find each of this μ and σ, EM algorithm is used and the results are substituted in Eq.(17).

As mentioned, EM algorithm has 4 steps. By substituting parameters defined in system model into EM algorithm, the Expectation and Maximization steps can be rewritten as:

E Step: Evaluating posterior probabilities:

$$P_{z,k} = \frac{\pi_k \mathcal{N}(t_z \mid \mu_k, \sigma_k)}{\sum_{k=1}^{K} \pi_k \mathcal{N}(t_z \mid \mu_k, \sigma_k)} \tag{17}$$

M Step: Re-estimating the parameters using computed parameters:

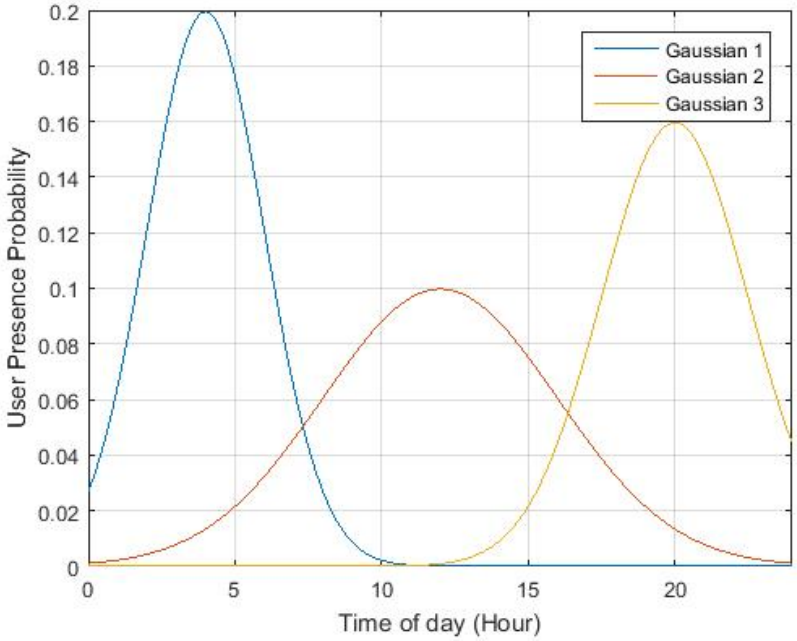

Figure 14: The smoothed user visit probability at different times during a day

$$\mu_k^{new} = \frac{1}{T}\sum_{z=1}^{T} P_{z,k} t_z \tag{18}$$

$$\sigma_k^{new} = \frac{1}{T}\sum_{z=1}^{T} P_{z,k}(t_z - \mu_k^{new})(t_z - \mu_k^{new})^T \tag{19}$$

$$\pi_k^{new} = \frac{T_k}{T} \tag{20}$$

where

$$T_k = \sum_{z=1}^{T} P_{z,k} \tag{21}$$

In the above formulas, each time slot t_z belongs to the triple $< u_i, l_j, t_z >$. T is the total number of time slots in one day. In this example, the accuracy is based on hour and T = 24, meaning that the location of the user is sampled at every 1 hour. T_k is the number of time slots for each of the Gaussian distribution components, it means T_k is equal to the time duration in which user presence at a specific location can modeled by an individual Gaussian distribution. In Figure 14 example, $T_k = 8$.

After EM algorithm calculations are done, the result is K new μ and σ, which are the estimated μ and σ for user predicted mobility pattern distribution at one specific location.

After doing all the above mentioned stages for each location, the presence probability of the user at each location l will be available.

3.2 User Total Available Bits Prediction

As it is explained above, predicting the amount of deliverable bits to the user at each time slot and location in the future, requires the predicted value of spectral efficiency and free radio resource at each time slot and location.

Multiplication of Spectral efficiency and free radio resource at each time slot and location results in total available bits to the user at that time slot and location. To predict these two parameters, linear prediction method is used. In the following subsections, spectral efficiency and free radio resource are explained briefly. The linear prediction is explained and used to predict the spectral efficiency and free radio resource.

3.2.1 Spectral Efficiency Prediction

Spectral efficiency which is measured by bits per second per hertz (bps/Hz), represents how efficient the radio spectrum is being used. Knowing spectral efficiency and available bandwidth, it is possible to calculate the available data rate to the user. In another words, spectral efficiency multiplied by the used bandwidth in a cell, is the cell throughput.

Item	Sub-category	LTE target	LTE-Advanced (4G) target	IMT-Advanced (4G) requirement
Peak spectral efficiency (b/s/Hz)	Downlink	16.3 (4x4 MIMO)	30 (up to 8x8 MIMO)	15 (4x4 MIMO)
	Uplink	4.32 (64QAM SISO)	15 (up to 4x4 MIMO)	6.75 (2x4 MIMO)
Downlink cell spectral efficiency b/s/Hz/user Microcellular 3 km/h, 500 m ISD	(2x2 MIMO)	1.69	2.4	
	(4x2 MIMO)	1.87	2.6	2.6
	(4x4 MIMO)	2.67	3.7	
Downlink cell-edge user spectral efficiency (b/s/Hz/user) (5 percentile, 10 users), 500m ISD	(2x2 MIMO)	0.05	0.07	
	(4x2 MIMO)	0.06	0.09	0.075
	(4x4 MIMO)	0.08	0.12	

Figure 15: Performance targets for LTE, Advanced-LTE, and IMT-Advanced [9].

Spectral efficiency can be influenced by different factors such as access technology, signal to noise ratio (SNR), modulation and coding scheme and number of antennas [29]. Figure 15 shows an example of spectral efficiency targets for 4G [9].

For predicting the spectral efficiency for a location in future, Linear Prediction method is used, which is explained in next subsection based on [30]. It should be explained that predicting spectral efficiency is a wide topic itself and can be done by many solutions whose detailed explanations are beyond the scope of this thesis. Therefore, the assumption here is that the historical value of spectral efficiency at each time slot and each location is measured and available.

As mentioned before, human mobility has a cyclic temporal continuity characteristic and as the users presence affects parameters such as SNR, it is assumed that the cyclic temporal continuity characteristic can be seen in spectral efficiency. Thus, it is possible to reconstruct the values of spectral efficiency at different locations during a day, based on the historical data which is supposed to be the samples of the spectral efficiency.

3.2.2 Linear Prediction

Early works on linear prediction theory was developed by Kolmogorov on the concepts of extrapolation of discrete time random processes. Further extensions of this work on multivariate processes were initiated by Levinson and Wiener. Nowadays, Linear Prediction is a fundamental theory in signal processing, speech processing and modeling of dynamic systems where it is used to predict the discrete time signal as a linear function of its past values. In this part, the goal is to discuss the equations and theory of linear prediction in detail.

Suppose $x(n)$ is wide sense stationary signal where the objective is to predict the value of $x(n)$ via a linear combination of N most recent samples. The common formulation of linear prediction has the form:

$$\widehat{x}_P(n) = \sum_{i=1}^{P} a_i x(n-i) \tag{22}$$

where p is a positive integer representing the prediction order, a_i denotes the prediction coefficients, and $\widehat{x}(n)$ is the estimated signal value. The prediction is defined as

$$e(n) = x(n) - \widehat{x}_P(n) \tag{23}$$

which represents the difference between true and estimated signal values.

The primary objective is to find the optimal prediction coefficients which can be obtained via minimizing mean square predication error ξ, expressed as:

$$\xi = E[\mid e(n) \mid^2] \tag{24}$$

Besides, the prediction error $e(n)$ can be redefined as the output response of a FIR filter $A(z)$ to input signal $x(n)$. The FIR transfer function is given by:

$$A(z) = 1 - \sum_{i=1}^{N} a_i z^{-i} \tag{25}$$

In order to reconstruct $x(n)$ from prediction error $e(n)$, the IIR filter $\frac{1}{A(z)}$ is utilized. Thus, linear prediction facilitates estimation of signal $x(n)$ from a set of prediction coefficients a_i. The optimal coefficients can be obtained through differentiating mean square error with respect to prediction coefficients a_i:

$$\frac{\partial \xi}{\partial a_i} = 0 \qquad i \leqslant i \leqslant P \tag{26}$$

which is equivalent to the fact that optimal coefficients are defined such that prediction error is orthogonal to $x(n-i)$.

$$E[e(n)x(n-i)] = 0 \qquad i \leqslant i \leqslant P \tag{27}$$

this can be summarized into normal equation:

$$\sum_{k=1}^{P} a_k r(i-k) = r(i) \tag{28}$$

where

$$r(i) = E[x(n)x(n-i)] \tag{29}$$

which denotes the autocorrelation of $x(n)$. The normal equations can be rewritten in matrix as

$$\underbrace{\begin{bmatrix} r(0) & r(1) & r(2) & \dots & r(p-1) \\ r(1) & r(0) & r(1) & \dots & r(p-2) \\ r(2) & r(1) & r(0) & \dots & r(p-3) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r(p-1) & r(p-2) & r(p-3) & \dots & r(0) \end{bmatrix}}_{R} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_p \end{bmatrix} = \underbrace{\begin{bmatrix} r(1) \\ r(2) \\ r(3) \\ \vdots \\ r(p) \end{bmatrix}}_{r}$$

or

$$Ra = r \tag{30}$$

Thereafter, the auto regression matrix R and r are formed, and prediction coefficients can be obtained using inverse matrix of autocorrelation multiplied by r. In addition, the prediction coefficients can be obtained without costly inversion of regression coefficients via Levinson-Durbin algorithm. The main condition to find optimal coefficients is non-singularity of autocorrelation matrix R. Moreover, the autocorrelation matrix R is a Toeplitz matrix. Also, the FIR filter $A(z)$ has minimum phase property where all its zeros z_k are inside the unit circle, leading to stable IIR filter $\frac{1}{A(z)}$.

Relying on the explanation of linear prediction, for constructing a predicted version of a series of values based on the samples, first, the prediction coefficients should

be found. This can be done by autocorrelation matrix $\widehat{S}_l(t_z)$ and using Wiener–Hopf equation $Ra = r$ [31].

Considering $\widehat{S}_l(t_z)$ as the spectral efficiency at location l at time t_z, and

$$r(i) = E[S_l(t_z) - S_l(t_{z-1})] \tag{31}$$

as the autocorrelation function, the prediction coefficients a_i can be found as the above matrix.

Using known prediction coefficients a_i and substituting the model parameters into the linear prediction formula, the predicted values of the spectral efficiency can be calculated by:

$$\widehat{S}_l(t_z) = \sum_{i=1}^{T_k} a_i S(t_{z-1}) \tag{32}$$

where $\widehat{S}_l(t_z)$ is the predicted value of spectral efficiency at location l and time t_z. T_k is the duration of the time in historical data, which is assumed as the prediction order.

Figure 16, shows an example of spectral efficiency at location l during one day and its predicted values using the prediction coefficients and linear prediction method.

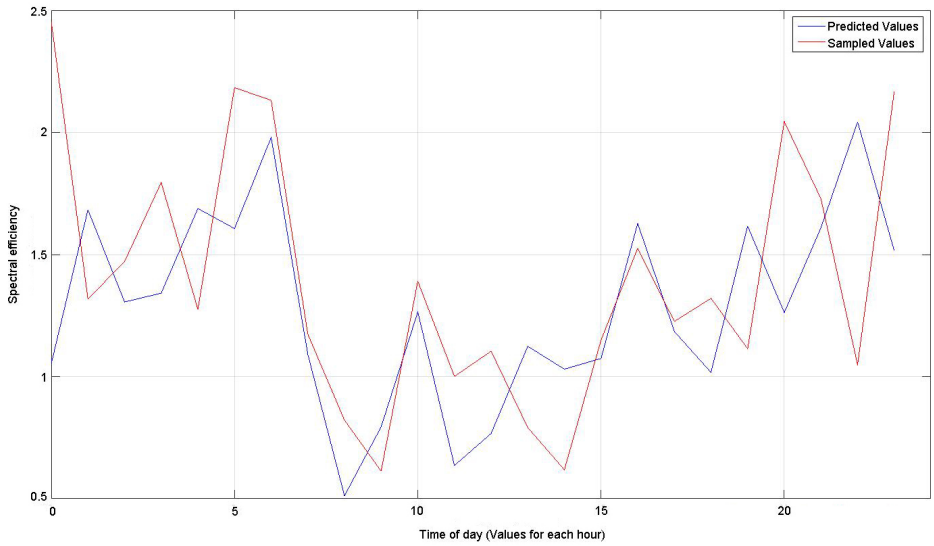

Figure 16: Historical spectral efficiency values and their predicted values during one day with accuracy of one hour

In Figure 16, the accuracy of prediction is one hour, means for a day the number of measured values is 24 and T_k, assumed as prediction order, is 24. By predicting the spectral efficiency for each 30 minutes instead of 1 hour, the accuracy of prediction

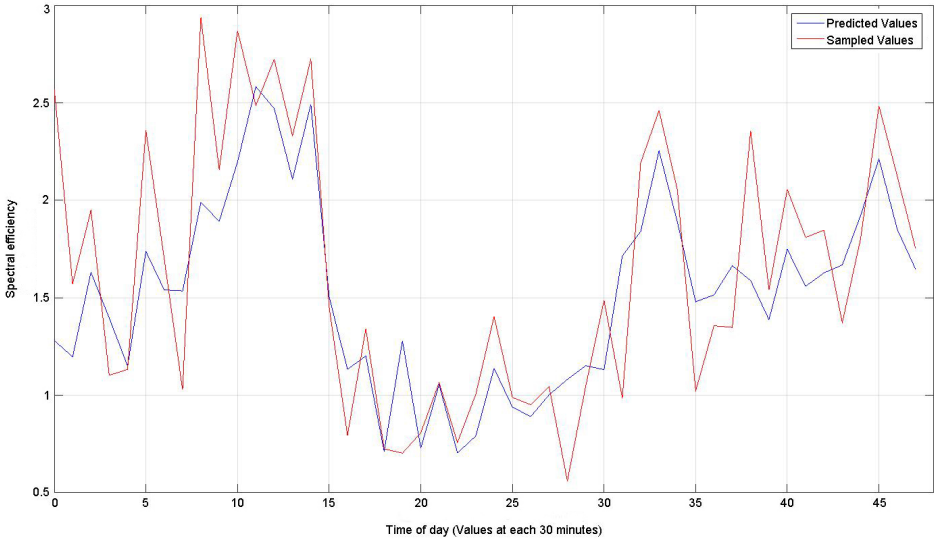

Figure 17: Historical spectral efficiency values and their predicted values during one day with accuracy 30 minutes

order changes from 24 to 48. As can be seen in Figure 17, which illustrates the spectral efficiency historical values and predicted values with accuracy of 30 minutes, the linear prediction can better follow the fluctuation of the samples values.

This becomes more clear when changing the prediction order to 288 corresponding to a 5-minute accuracy. Figure 18 shows spectral efficiency sample values and their predicted values during a day with the accuracy of 5 minutes.

3.2.3 Available Resource Prediction

Available free resource to the user in this model is radio resource which is measured in seconds times hertz ($seconds * hertz$). Figure 19 shows an example of LTE resource block.

The prediction method used in this model for available resource at a specific time and location in the future is Linear Prediction which is the same method applied for prediction of spectral efficiency.

By considering $\widehat{R}_l(t_z)$ as the available resource at location l at time t_z, $r(i) = E[R_l(t_z) - R_l(t_{z-1})]$ is the autocorrelation function and the prediction coefficients a_i can be found through the mentioned matrix.

By substituting the model parameters and calculated prediction coefficients a_i into the linear prediction formula, the predicted values of the spectral efficiency can be calculated by:

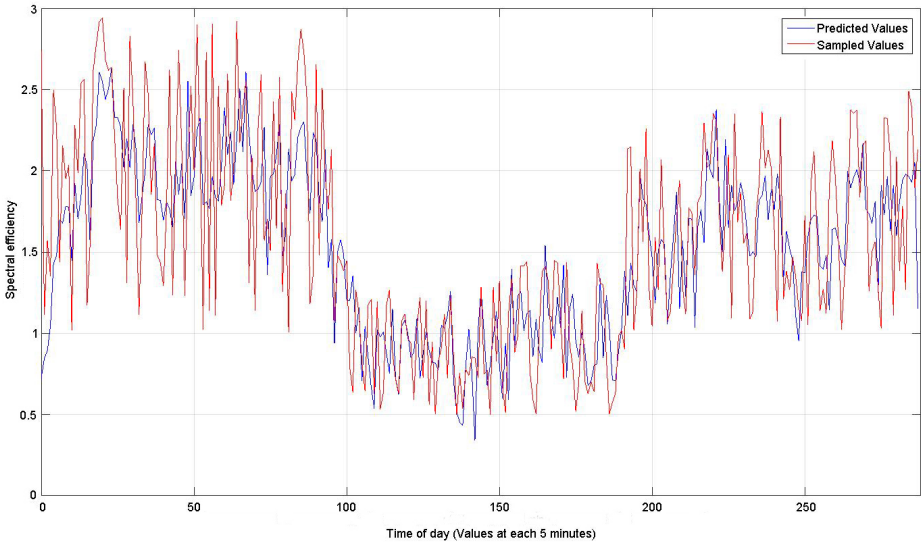

Figure 18: Historical spectral efficiency values and their predicted values during one day with the accuracy 5 minutes

$$\widehat{R}_l(t_z) = \sum_{i=1}^{T_k} a_i R(t_{z-i}) \tag{33}$$

where $\widehat{R}_l(t_z)$ is the predicted value of spectral efficiency at location l and time t_z, and T_k is the duration of the time in historical data, (taken as the prediction order).

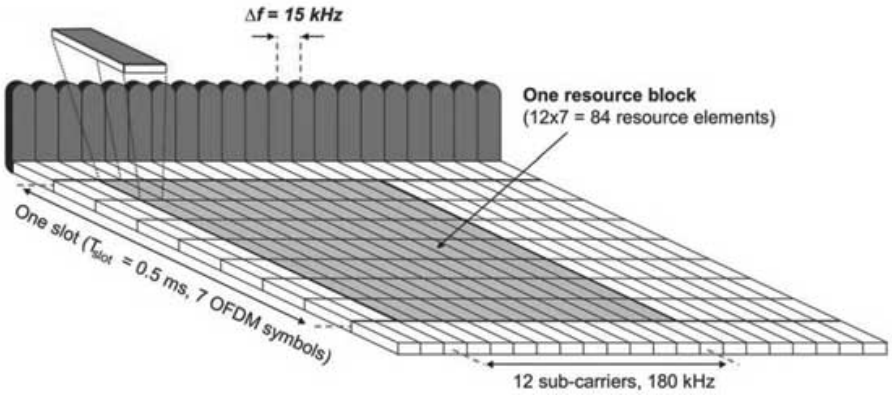

Figure 19: LTE downlink physical resource in a time-frequency grid [10].

Figure 20 and 21 show examples of historical radio resource samples with their predicted values using linear prediction. Figure 20 has an accuracy of one hour with 24 values and Figure 21 has an accuracy of 5 minutes with 288 values.

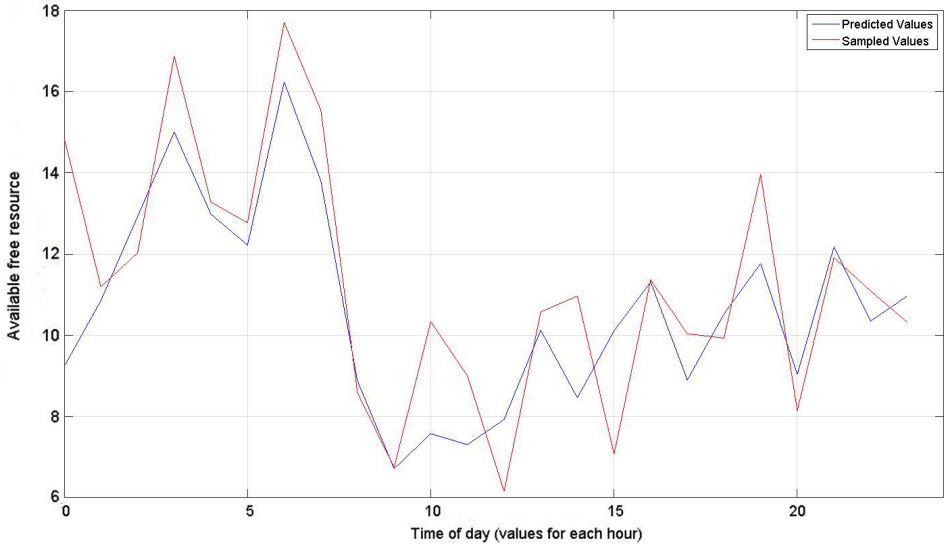

Figure 20: Historical free available resource values and their predicted values during one day with an accuracy of hour.

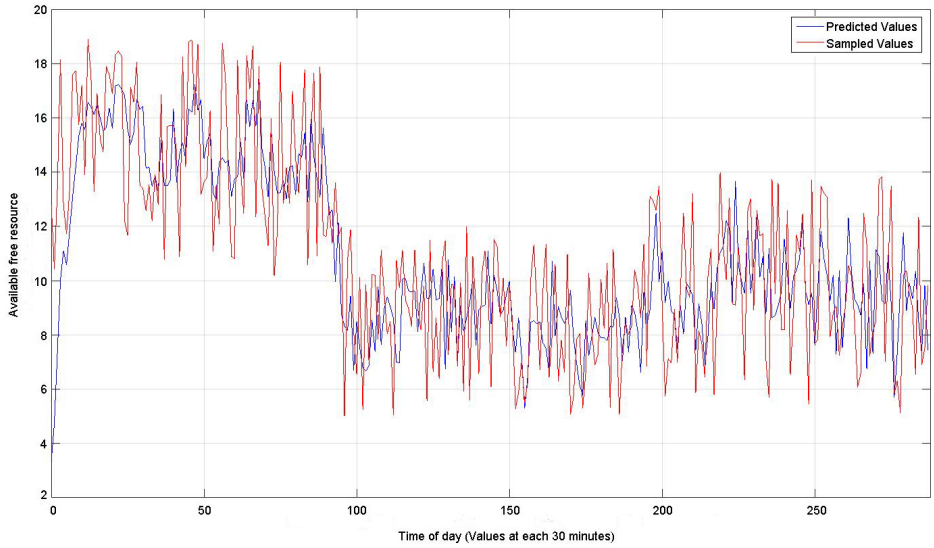

Figure 21: Historical free available resource values and their predicted values during one day with the accuracy of 5 minutes.

3.2.4 Total Available Bits to User

For calculating the total predicted bits that can be available to the user at location l and time t_z, spectral efficiency and available free resource at location l and time t_z, should be multiplied together.

$$B_l(t_z) = S_l(t_z) \times R_l(t_z) \tag{34}$$

where $B_l(t_z)$ is the total available bits to the user at location l at time t_z. Total available bits unit is bit as:

$$(\frac{\frac{bit}{sec}}{hz}) \times (sec * hz) = (bit)$$

Figure 22 shows an example of total available bits to the user at location l during one day with accuracy of one hour. This figure is the multiplication of values in Figure 16 and Figure 20.

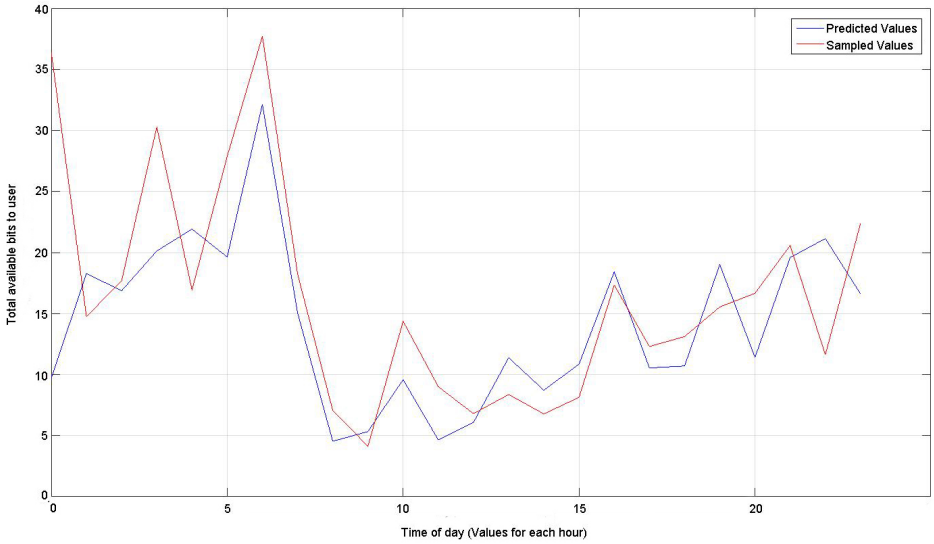

Figure 22: Total available bits to user at different times during a day, calculated by historical data and its predicted values.

3.3 User Available Bits Probability

At this stage, user presence probability and total received bits available to the user at specific time t_z and location l in the future is predicted already. Now it is possible to calculate the probability of the total available bits to the user at any time t_z. It can be considered as PMF (Probability Mass Function) of total available bits to the user at any time t_z.

It should be noted that as the total received bits value at any time t_z is a countable set, since the number of variables is equal to the number of locations, PMF should be used instead of PDF.

General form of probability mass function of discrete random variable X with countable range of $R_x = x_1, x_2, x_3, ..., X_n$, is in the function-form below:

$$P_x(x_k) = P(X = x_k) \qquad k = 1, 2, 3, ..., n \tag{35}$$

Two main properties of mass functions are [32]:

$$0 \leqslant P_x(x_k) \leqslant 1 \qquad \text{for all } x$$

$$\sum_{x_k \in R_x} P(x_k) = 1$$

For calculating the PMF of total available bits to the user at time t_z, Eq.(33) can be rewritten as:

$$P_{B_l(t_z)}(L_j) = Pr(L = L_j) \tag{36}$$

where L_j indicates the probability of the user being at location j.

Figure 23 illustrates an example of a PMF diagram of discrete distribution function $X1$ at time t_z, which is assumed as the distribution function of total available bits to the user at time t_z. If the X and Y axes assume the amount of total available bits to the user and their probabilities respectively, Figure 23 can be considered as PMF of total available bits to the user at time t_z regardless of user's location.

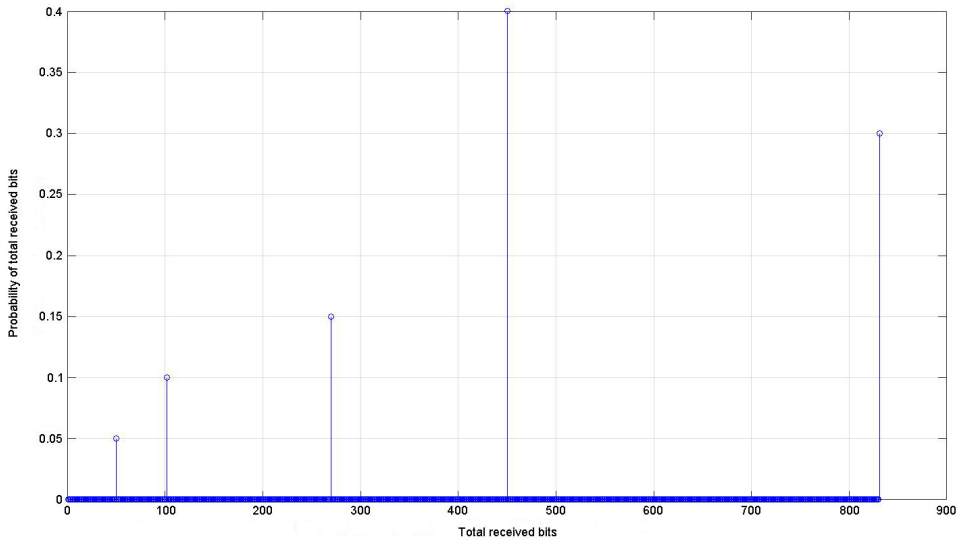

Figure 23: PMF for descret variable series $X1$ at time t_z

Using the PMF of time t_z, it is possible to create the CDF of total available bits to the user at time t_z, which is shown in Figure 24.

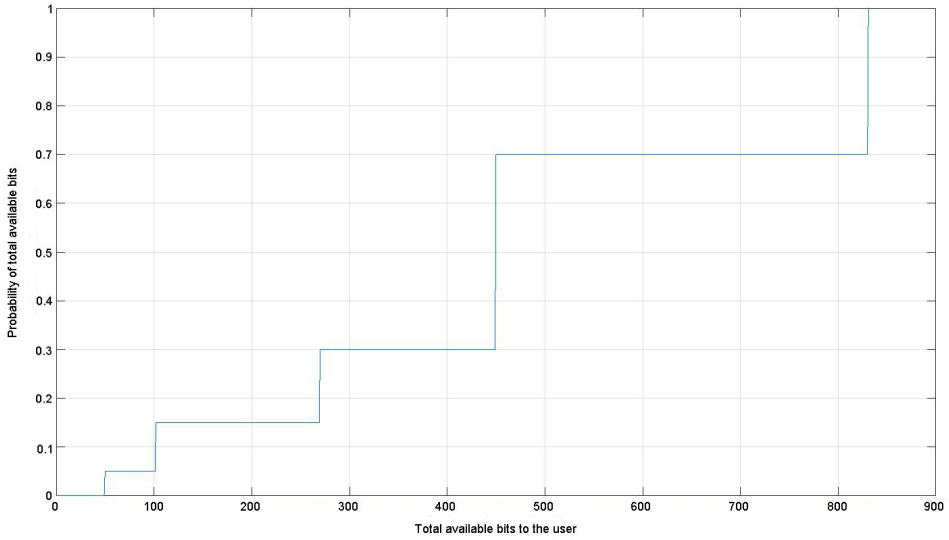

Figure 24: CDF of total available bits to the user at time t_z

Based on the above description, for each time slot t_z, there can be a PMF of total received bits representing the probability of total amount of bits deliverable to the user at each location. Usually the result for only one time slot is not useful rather it is required that the total amount of bits that can be delivered to the user during a specific time duration be calculated.

For calculating the total available bits to the user at a duration of time $t_k = \sum_{i=1}^{K} t_i$, which is called time window in this model, it is necessary to find the distribution function of discrete variable t_k. As t_k is the sum of two or more independent discrete functions, the convolution of those functions is the distribution function of t_k [33].

If one assumes $A(x)$ and $B(x)$ as distribution functions of independent random variables X and Y, $Z = X + Y$ and $C(x)$ be the distribution function of Z, then

$$P(Z = z) = \sum_{j=-\infty}^{\infty} P(X = k)P(Y = z - j) \tag{37}$$

and the distribution function of $C(x) = A(x) \times B(x)$ which is the convolution of $A(x)$ and $B(x)$ is [33]:

$$C(x) = \sum_{(j)} A(j)B(x - j) \tag{38}$$

where j is represents the time duration in which convolution is done.

By substituting the model parameter into Eq.(36), and considering m_j as the distribution function of total available bits probability to the user at time $t(j)$ the

distribution function of $t_k = \sum_{i=1}^{2} t_i$ can be written as:

$$t_k(x) = \sum_{B} m_1(B) m_2(x - B) \tag{39}$$

If t_k contains more than two independent discrete functions, the convolution result of the first two functions should be convolved with the third function and so on till the last time slot.

Figure 25 illustrates an example of a PMF diagram of discrete distribution function $X2$, which is assumed as the distribution function of total available bits to the user at time t_{z+1}.

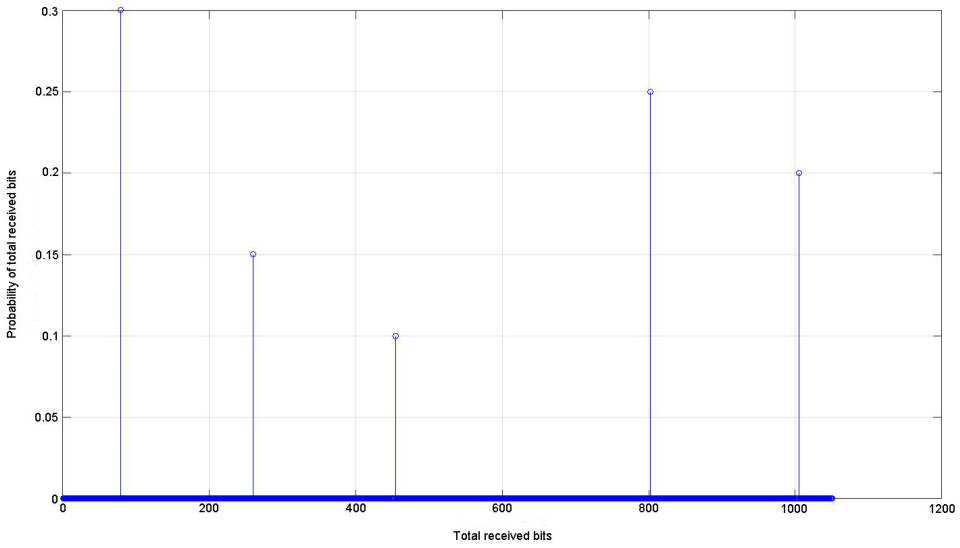

Figure 25: PMF for descret variable series $X2$ at time t_{z+1}

Figure 26 shows an example of convolution of distribution functions $X1$ and $X2$ (plotted in Figure 23 and Figure 25). In fact, Figure 26 presents the distribution function of total available bits to the user at time window $t_k = t_z + t_{z+1}$.

Now it is possible to find the CDF of total bits received by the user during time window t_k, using the distribution function of t_k. Figure 27 shows an example of CDF of total bits received by the user which is plotted as PMF in Figure 26.

Comparing the CDF of total available bits to the user at time t_z and during time window $t_k = t_z + t_{z+1}$ in Figure 28, shows that the delivery probability of a specific number of bits to the user in time window t_k, is less than time slot t_z. This does not represents a good interpretation of the model as the delivery probability of a specific number of bits to the user in time window t_k should be more than time slot t_z. For having a better interpretation for the model, complementary cumulative distribution

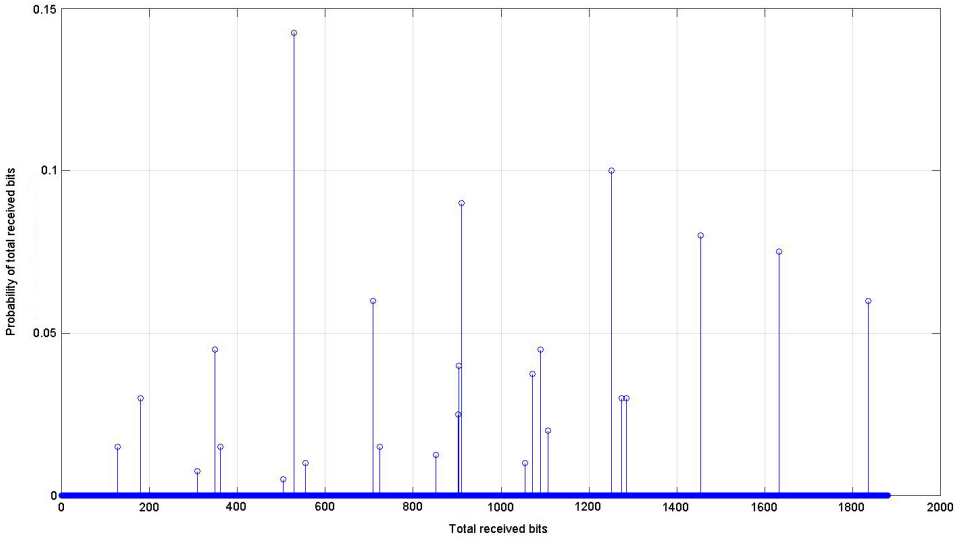

Figure 26: Convolution of two discrete probability distribution of $X1$ and $X2$ at time window $t_k = t_z + t_{z+1}$

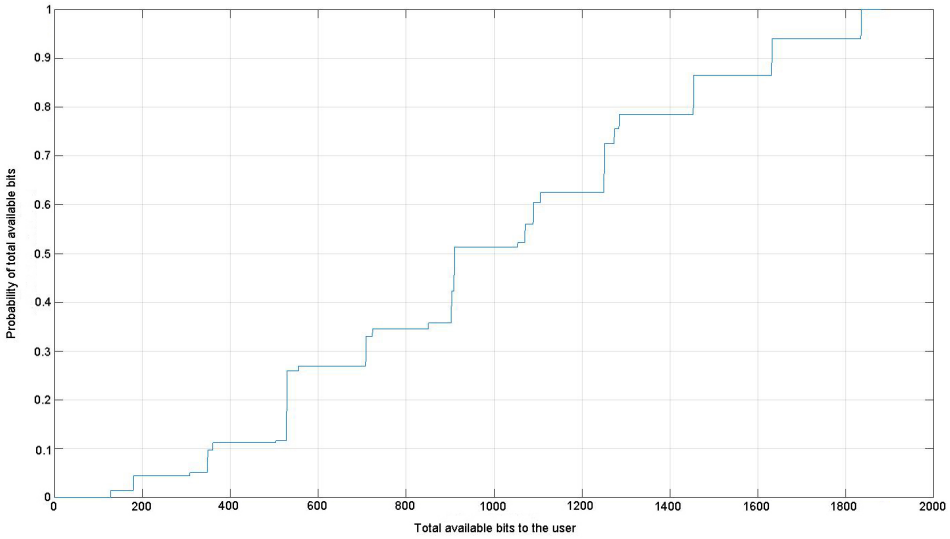

Figure 27: CDF of total received bits to user time window $t_k = t_z + t_{z+1}$

function (CCDF) should be used.

CCDF, also known as survival function [34], shows the survival of the system beyond a specific time and can be calculated by complementing the CDF which is 1-CDF.

Figure 29 shows the CCDF of total available bits to the user at time t_z and during time window $t_k = t_z + t_{z+1}$.

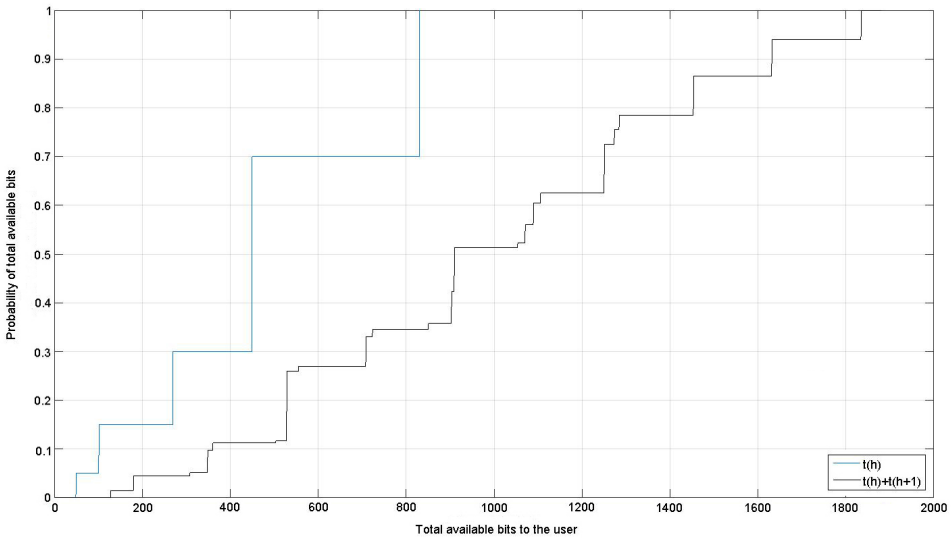

Figure 28: CDF of total received bits to user at time t_z and time window $t_k = t_z + t_{z+1}$

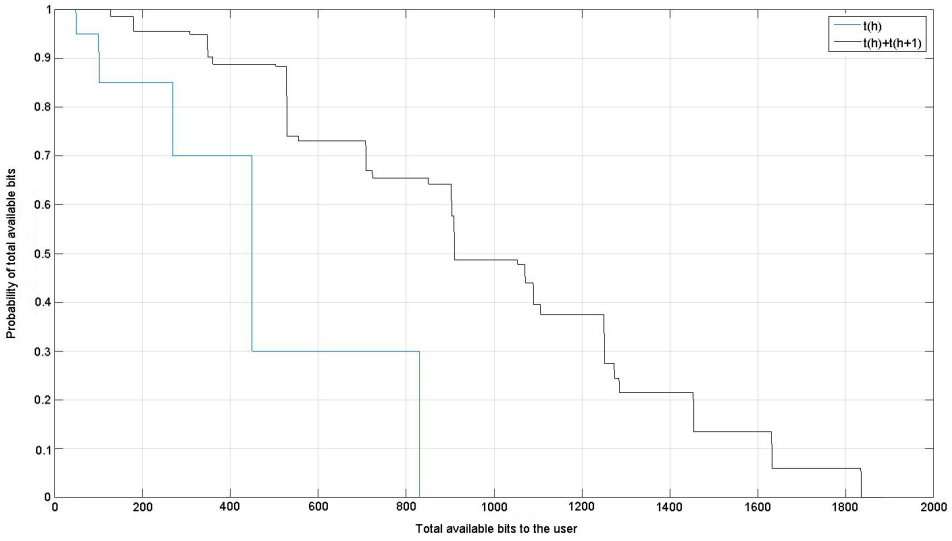

Figure 29: CCDF of total available bits to user at time t_z and time window $t_k = t_z + t_{z+1}$

4 Simulation Model

In the previous chapter a framework model was suggested. In this chapter, the suggested model is used to predict the probability of the required time to transfer a specific amount of data to the user in future, relying on a real life scenario for a user. The presumed real life scenario in this model is based on the routine daily route of a student or an employee.

As there was no access to real historical data, in this thesis the historical data are generated based on a real life scenario. For simplicity, the general assumption in this model and all the simulations is that the network is going to serve only one user and its behavior is separated into 3 main parts during a day based on a real life scenario of a normal employee or student. The user is usually at home during 00:00 to 08:00 and at work or university during 08:00 to 16:00. From 16:00 to 00:00 it is more possible that the user be at home or around city for different events or activity, than at work.

As a balance between simplicity and good accuracy, the presence probabilities, spectral efficiency and available free resource are considered to be fixed and roughly the same for each 5 minutes. It means the samples and estimations are done for each 5 minutes. Therefore, $t(h) = t_z = \{5, 10, 15, ...\}$ where $z = \{1, 2, 3, ...\}$.

4.1 Simulating Predicted User Location

Based on the scenario, the network has 2 main cells, and each cell is called a location and thus $L = \{l_1, l_2\}$. Whenever the user is not at any of these locations, it is considered that the user is out of range and cannot be served by the network. Location l_1 is considered as a residential place and most of the buildings in that location are homes and apartments. Location l_2 is considered as the administrative section which mostly consists of organizational buildings, companies, university, and similar buildings. Hence, at each time t_z the user location has only three probabilities, being at location l_1, location l_2 or out of network range.

Considering the facts and conclusion in Section 3.1, user presence during a day at a specific location follows a mix of a few Gaussian distributions and it is modeled by GMM. To find the proper μ and σ to create the GMM, the EM algorithm is used which requires initial μ and σ for each Gaussian distribution. These initial μ and σ should be selected from the historical presence probability of the user. Different techniques can be used to find these initial μ and σ, for example, for each period of time, the time stamp which has the highest presence probability can be used as the initial μ . The initial σ can be considered as the variation in the range of visiting frequency before and after the time stamp which is considered as the initial μ.

As mentioned in the beginning of this chapter, the daily mobility pattern of the user consists of 3 main sub-patterns based on the time of day. The assumption in

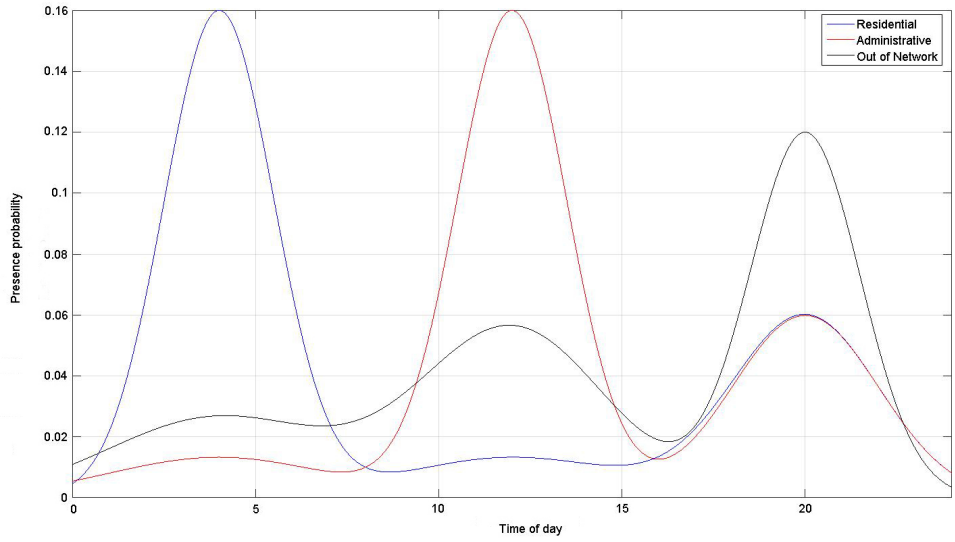

Figure 30: User presence probability during one day at different locations based on the historical data

this scenario is that the center time stamp of each of these sub-patterns can be used as the initial μ, and the initial σ is based on the variation in the range of visiting frequency before and after the initial μ.

Figure 30 shows the user historical presence probability during a day at locations l_1, l_2 and being of out network range.

As can be seen in Figure 30, during a day, for each location, three main Gaussian distributions can be recognized with μ at 4, 12 and 20. They can be used as the initial μ values for each location. Based on the checking frequency of the user around these μ values, three approximate values can be selected randomly for σ.

By assuming

$$\mu_{1,1} = 4, \quad \sigma_{1,1} = 1.5 \qquad \mu_{1,2} = 12, \quad \sigma_{1,2} = 3 \qquad \mu_{1,3} = 20, \quad \sigma_{1,3} = 2$$
$$\mu_{2,1} = 4, \quad \sigma_{2,1} = 3 \qquad \mu_{2,2} = 12, \quad \sigma_{2,2} = 1.5 \qquad \mu_{2,3} = 20, \quad \sigma_{3,3} = 2$$
$$\mu_{3,1} = 4, \quad \sigma_{3,1} = 3 \qquad \mu_{3,2} = 12, \quad \sigma_{3,2} = 2.5 \qquad \mu_{3,3} = 20, \quad \sigma_{3,3} = 1.5$$

as the initial μ and σ values for each Gaussian at each location and using EM algorithm, Equations 27 to 31, the estimated μ and σ values are:

$$\mu_{1,1}^{new} = 3.98, \quad \sigma_{1,1}^{new} = 1.37 \qquad \mu_{1,2}^{new} = 12.5, \quad \sigma_{1,2}^{new} = 2.87$$

$$\mu_{1,3}^{new} = 19.34, \quad \sigma_{1,3}^{new} = 2.3$$

$$\mu_{2,1}^{new} = 3.92, \quad \sigma_{2,1}^{new} = 3.08 \quad \mu_{2,2}^{new} = 12.08, \quad \sigma_{2,2}^{new} = 1.49$$
$$\mu_{2,3}^{new} = 20.08, \quad \sigma_{3,3}^{new} = 1.63$$

$$\mu_{3,1}^{new} = 3.82, \quad \sigma_{3,1}^{new} = 3.4 \quad \mu_{3,2}^{new} = 12.27, \quad \sigma_{3,2}^{new} = 2.5$$
$$\mu_{3,3}^{new} = 20.11, \quad \sigma_{3,3}^{new} = 1.33$$

where each $\mu_{i,j}$ and $\sigma_{i,j}$ belong to j^{th} Gaussian at location i.

Comparing the initial and estimated values for each $\mu_{i,j}$ and $\sigma_{i,j}$, proves that the EM algorithm output is accurate enough for estimating the presence probability of the user.

Figures 31, 32 and 33 represent the user presence probability during a day based on the historical data and the estimated values at each location.

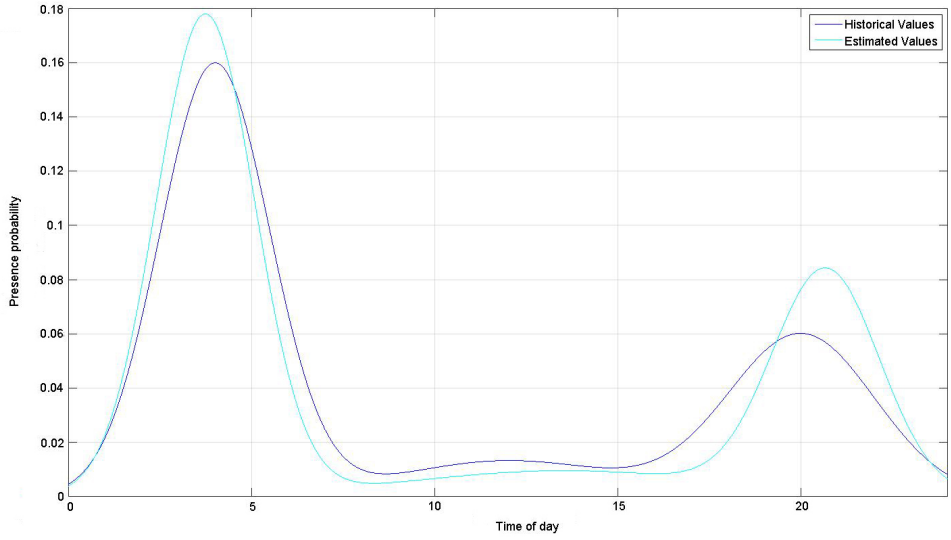

Figure 31: User presence probability during one day at l_1 based on the historical data and its estimated values.

The last step for this part is normalizing the presence probability of the user for each t_z. Figure 34 shows the normalized user presence probability at each location.

4.2 Simulating Predicted Spectral Efficiency

The assumption about time and locations in the previous part are considered for Spectral Efficiency prediction. The additional assumption here is that location l_2, as administrative area, provides higher spectral efficiency than l_1, as residential area,

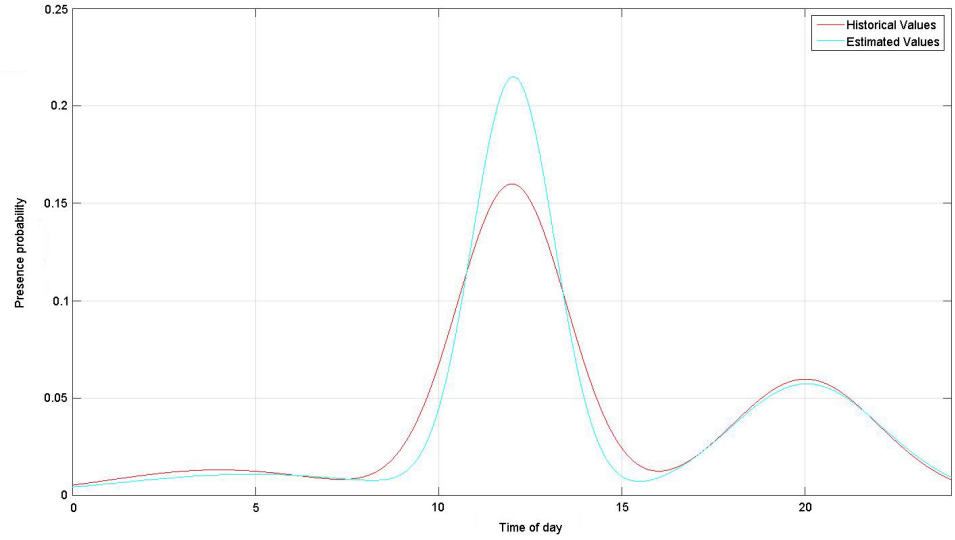

Figure 32: User presence probability during one day at l_2 based on the historical data and its estimated values.

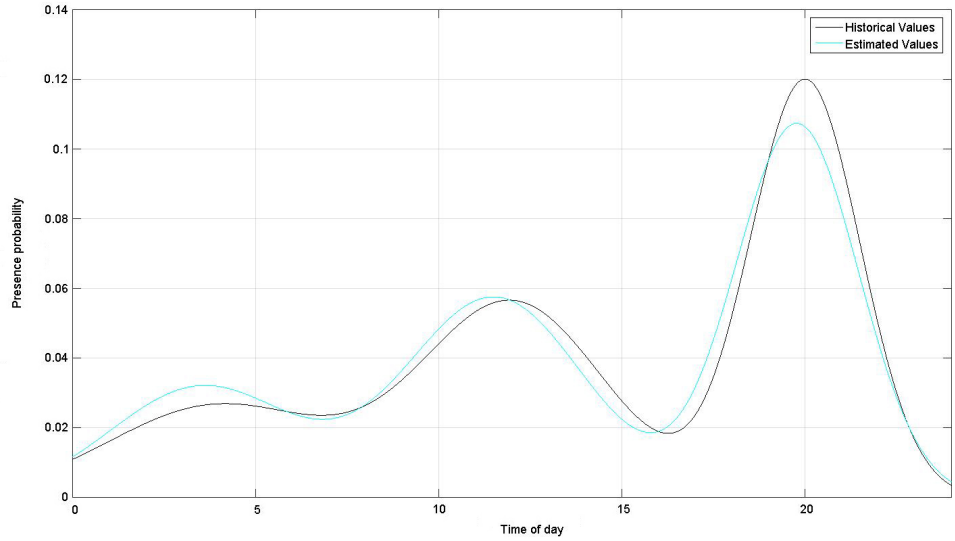

Figure 33: User being ou of range probability during one day based on the historical data and its estimated values.

and it results in higher data rate to the user.

As it was suggested in Section 3.2, the linear prediction method is used for prediction of spectral efficiency values in the future by sampling from network historical spectral efficiency statistics.

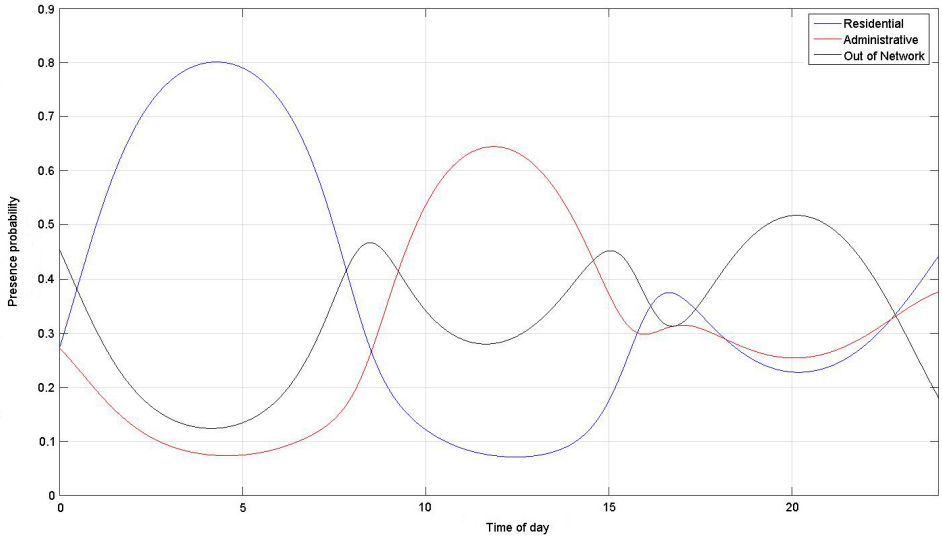

Figure 34: Normalized user presence probability for for each t_z

by substituting simulation model values into Eq.(33), it can be rewritten as:

$$\widehat{S_l}(t_z) = \sum_{i=1}^{864} a_i S(t_{z-1})$$

where $T_k = 864$, as 24 hours includes 288 five-minute intervals, and for more accuracy, historical data for 3 days is used for sampling. Prediction coefficients a_i can be found by using the autocorrelation function in Eq.(32)and auto regression matrix R.

Figures 35, 36, and 37 illustrate the historical spectral efficiency values of network N and the predicted values at locations l_2 for the 3 different days in a row used for sampling.

By comparing the predicted values and the historical values for each day, it can be seen that the linear prediction method with 3 days sampling has the proper accuracy for model simulation. although, more accurate predictions are achieved when using more number of days for sampling. The proper accuracy here is estimated by simulating the prediction with different sampling numbers, and 3 days are chosen as a good trade off between complexity and accuracy. Based on the data types and usage of linear prediction, the proper sampling number may be vary.

For location l_2, the same method had been used and Figure 38, shows the third sample day values and the predicted values for spectral efficiency.

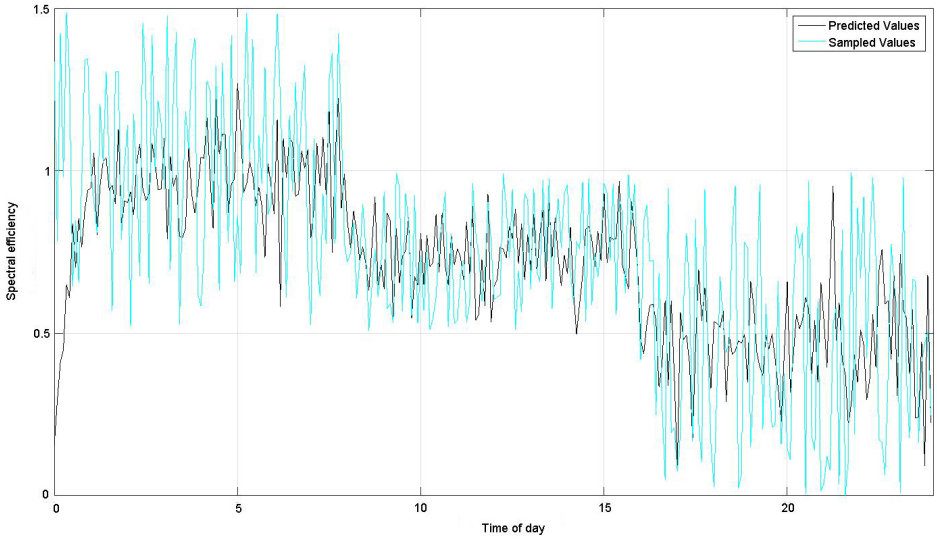

Figure 35: Historical spectral efficiency values for the first day and predicted spectral efficiency values based on 3 days for location l_1

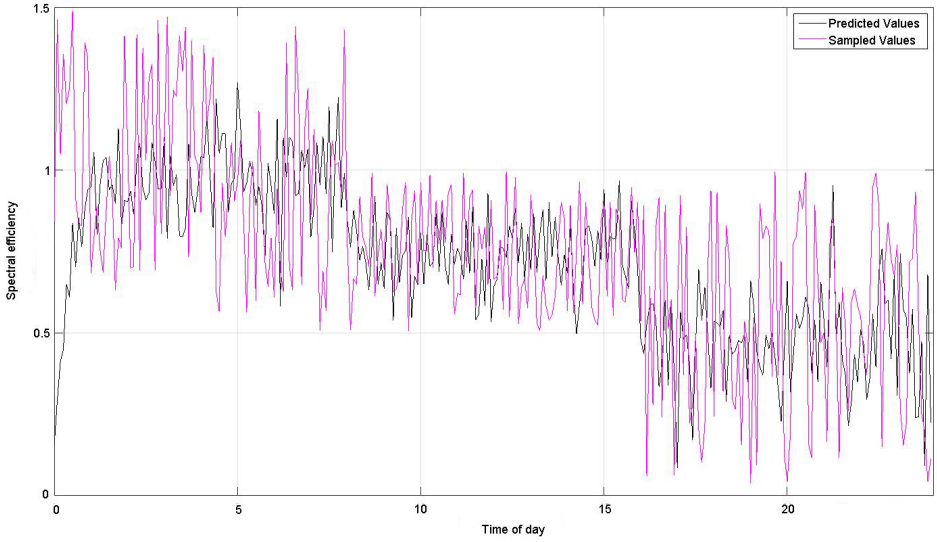

Figure 36: Historical spectral efficiency values for the second day and predicted spectral efficiency values based on 3 days for location l_1

4.3 Simulating Predicted Available Resource

The entire assumption and methodology for predicting the radio resource are the same as spectral efficiency. By substituting simulation model values into Eq.(34), it can be rewritten as:

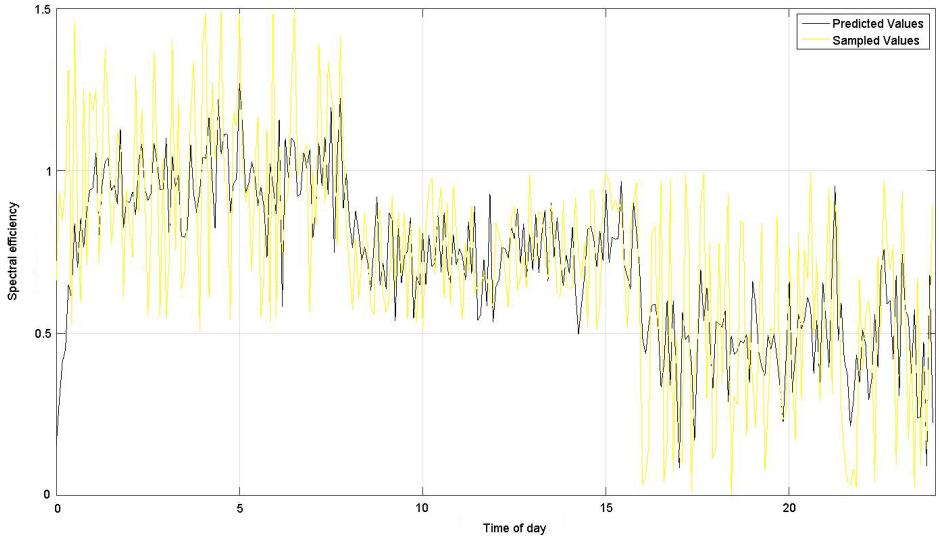

Figure 37: Historical spectral efficiency values for the third day and predicted spectral efficiency values based on 3 days for location l_1

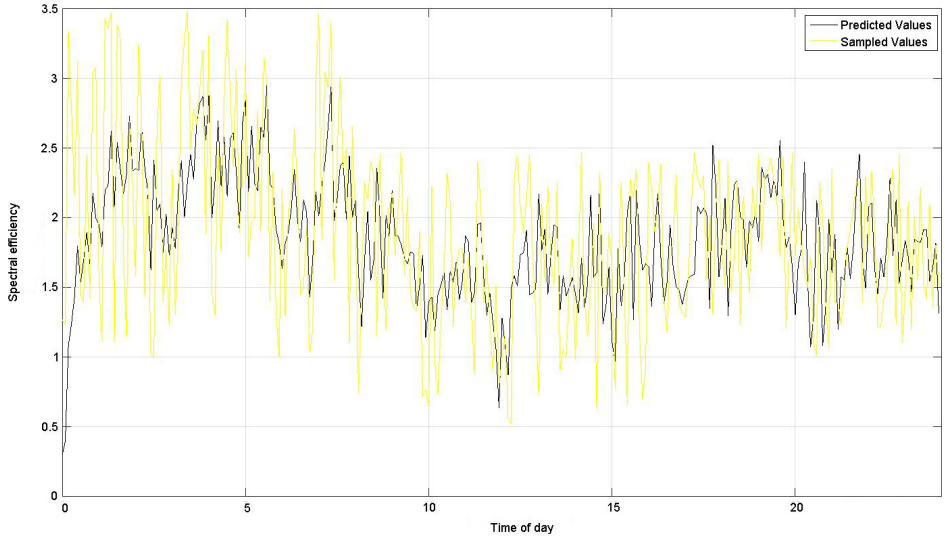

Figure 38: Historical spectral efficiency values for the third day and predicted spectral efficiency values based on 3 days for location l_2

$$\widehat{R_l}(t_z) = \sum_{i=1}^{864} a_i R(t_{z-1})$$

Figure 39 and 40 belong to historical and predicted available resource at locations l_1 and l_2 respectively.

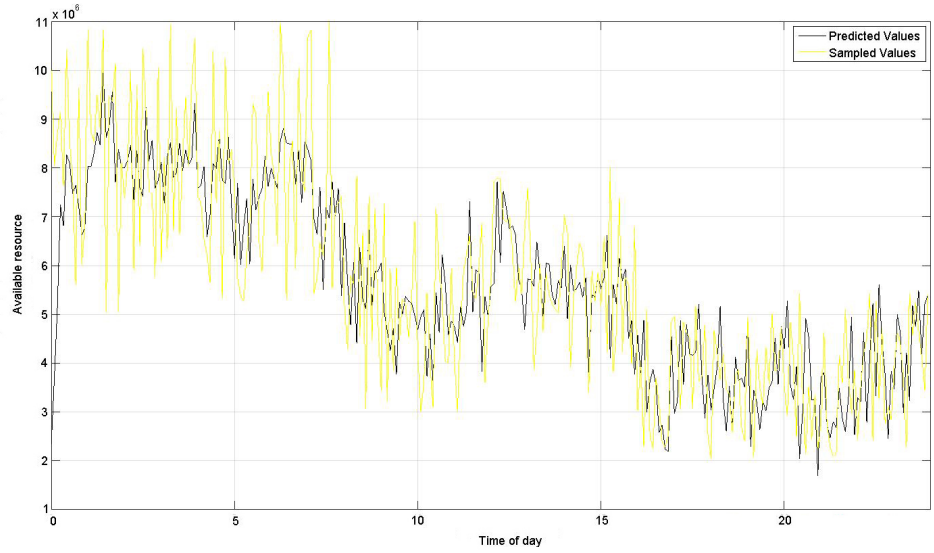

Figure 39: Historical available free resource values for the third day and predicted available free resource values based on 3 days for location l_1

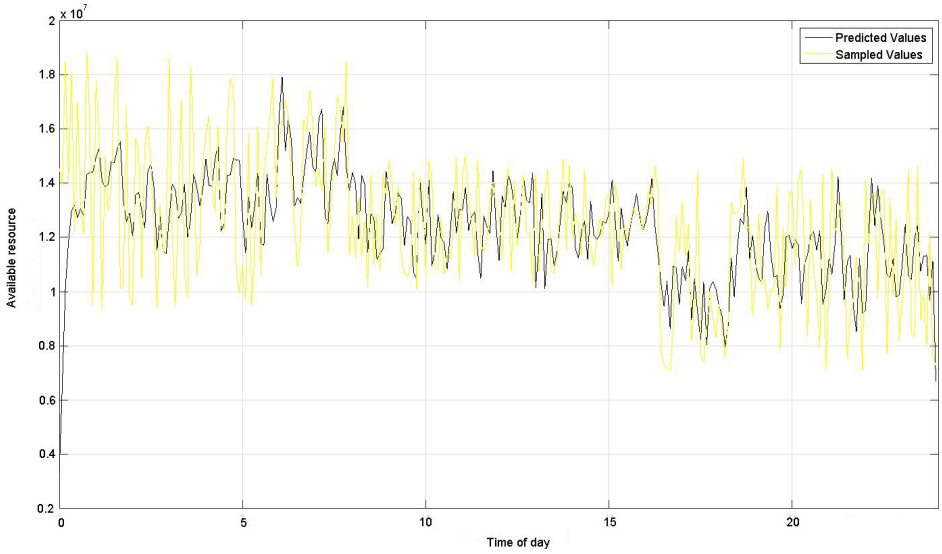

Figure 40: Historical available free resource values for the third day and predicted available free resource values based on 3 days for location l_2

5 Experimental Results

In Chapter 4, all the required parameters and values for calculating the probability of the total available bits to the user in the future were estimated. In this chapter, the probability of the total available bits to the user in future and the probability of the required time to deliver a specific amount of data to the user is calculated.

5.1 Predicted Total Available Bits

Following the described methodology in the Chapter 4, the total available bits the user at each time and location is the multiplication of the free resource and spectral efficiency at that time and location.

Until now, the predicted spectral efficiency and available resource for different times t_z and different locations l_1 and l_2 are computed. Figure 41 illustrates the total available bits to the user at locations l_1 and l_2 based on Eq.(35), where $l = \{1, 2\}$ and $l = \{1, 2\}$

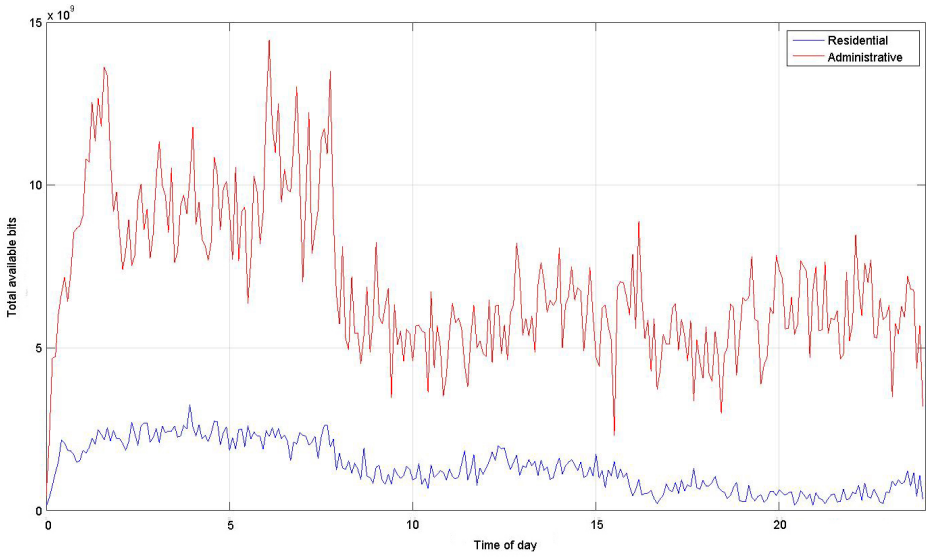

Figure 41: Predicted total available bits to the user during a day.

5.2 Probability of Total Available Bits

Based on the explanation of the previous chapter, it is possible to create a PMF of total available bits to the user for each time t_z. Figure 42, shows an example of PMF for time $t(120)$ indicating the time interval 10:00-10:05, considering the assumption that $t_1 = 00{:}00\text{-}00{:}05$ and each t_z is equal to 5 minutes. Figure 43 shows the PMF for time $t(121)$ corresponding to 10:05-10:10.

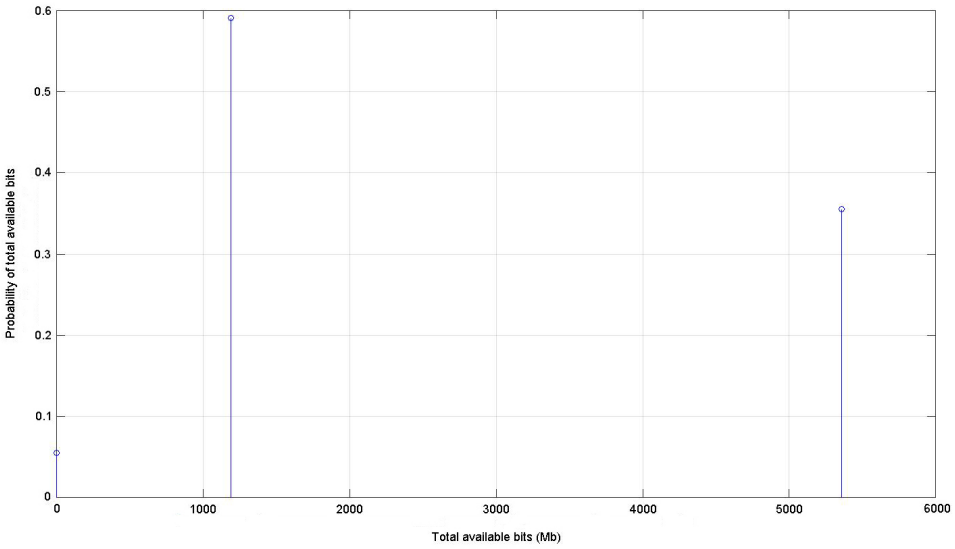

Figure 42: PMF of total available bits to user at time 10:00 to 10:05

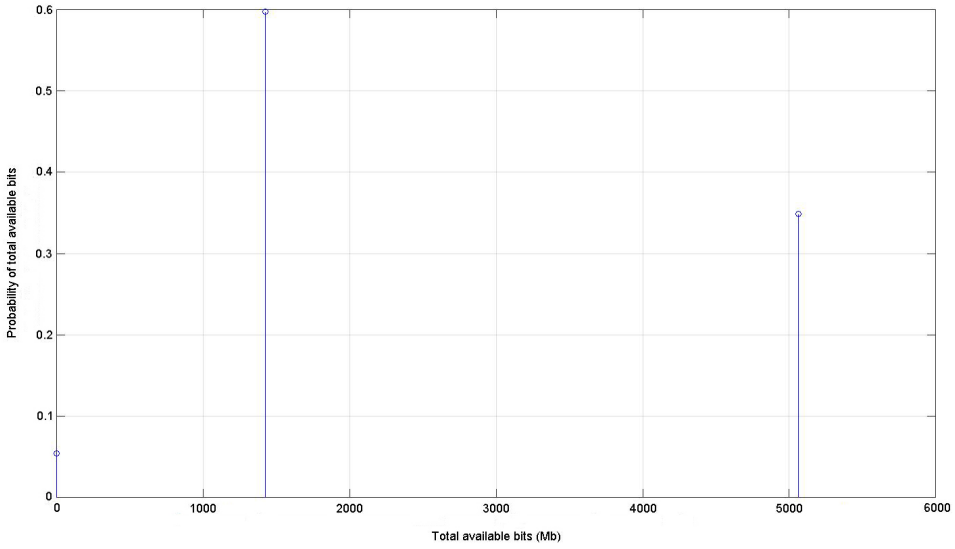

Figure 43: PMF of total available bits to user at time 10:05 to 10:10

By convolving these two PMF, the PMF of total available bits to the user during a 10-minute time window, $T_k = 10$, can be calculated. Convolving these PMF for each time slot t_z, makes it possible to find the CCDF of total available bits during different time windows.

Figure 44 represents the CCDF of total available bits to the user during $T_k = 30$ min to $T_k = 240$ min as an example.

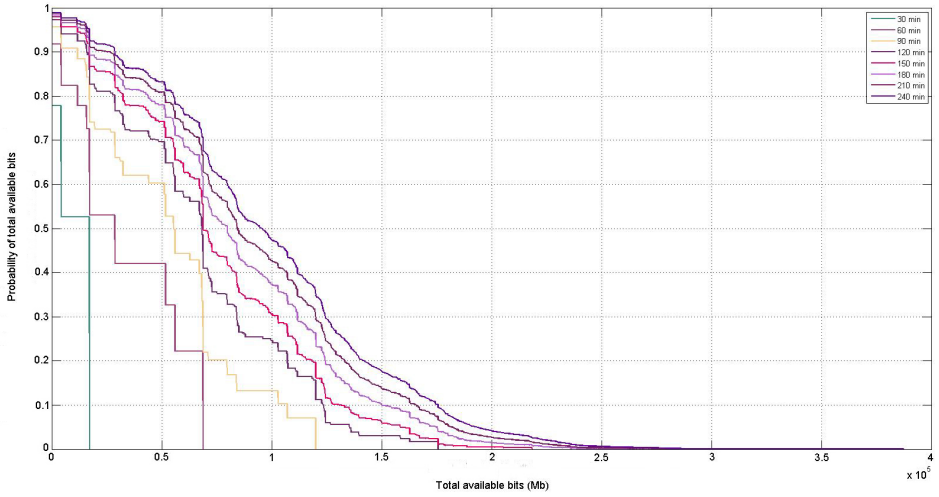

Figure 44: Probability of total available bits to the user during different times windows from 00:00 to 04:00

The presented CCDFs in Figure 44 belong to a 4-hour time duration divided to 30-minute to 240-minute time windows starting at 00:00. Figure 45 shows the CCDF of total available bits to the user in time windows with the same durations but it starts at 14:00.

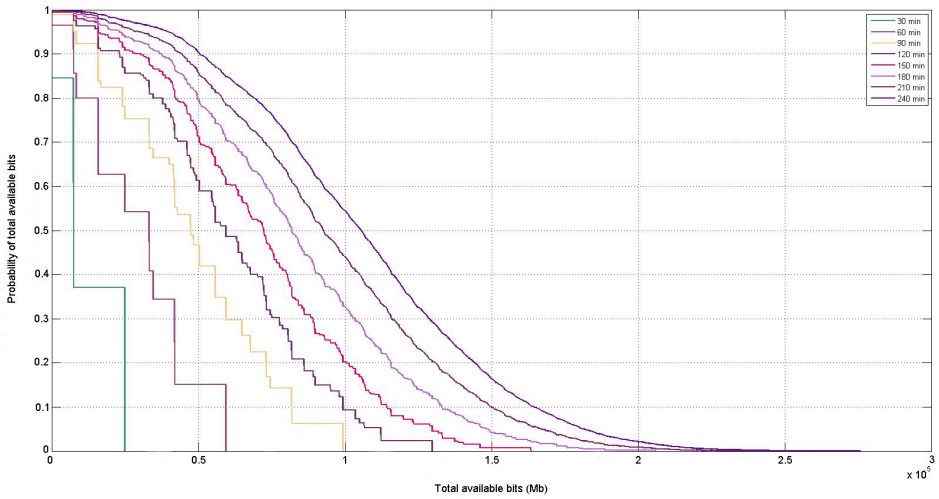

Figure 45: Probability of total available bits to the user during different times windows from 14:00 to 18:00

Comparing Figures 44 and 45 shows that the starting time of delivering data has a significant effect on the probability of total bits that can be delivered to the user. For example, the probability of delivering 20 GB of data to the user during 180 minutes, is about 20% if it starts at 14:00 and it is around 60.5% if it starts at 00:00.

5.3 Delivery Time Prediction

In this chapter, 3 sample files with volumes of 4.5, 9 and 22.5 GB, which are similar to a 2-hour HD movie, 4-hour extended version of a HD movie and a 2-hour UHD movie are given to the system to predict the probability of the time duration required to transfer this amount of data to the user.

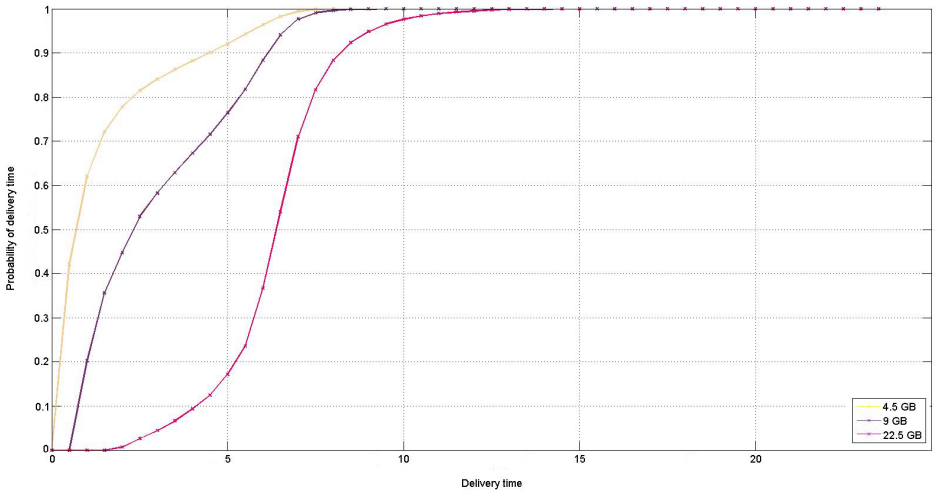

Figure 46: Delivery time probability for 4.5, 9 and 22.5 *GB* of data, starts at 00:00

Figures 46 and 47, presents the delivery time probability if starting at 00:00 and 10:00, respectively.

It can be seen that the system is capable of transferring a 4.5 GB HD movie by 100% in 9 hours if delivery starts 00:00 and in 13.5 hours to transfer 22.5 GB UHD movie by 100%. Transferring a UHD movie in 3 hours is risky as its delivery probability is 37%.

Another type of analysis that can be done based on Figure 46 is finding the duration of time for transferring a specific amount of data with a specific probability. For example the minimum time duration for transferring a HD movie with the probability of equal to or more than 90% is 4.5 hours.

Comparing Figure 46 with Figure 47, reveals the effect of delivery start time. As it is shown in Figure 47, the system is capable of transferring a UHD movie in 14 hours by 100% which is 30 minutes more than the delivery start time of 00:00. Another comparison shows that an extended HD movie can be delivered to the user during 6.5 hours with the probability of 90% if the delivery starts at 00:00 and it takes 4.5 if delivery starts at 10:00 which is 44% faster.

In fact, the effect of delivery start time is the results of users presence probability

at different locations at different times and the total available bits to the user at those locations during the user presence there.

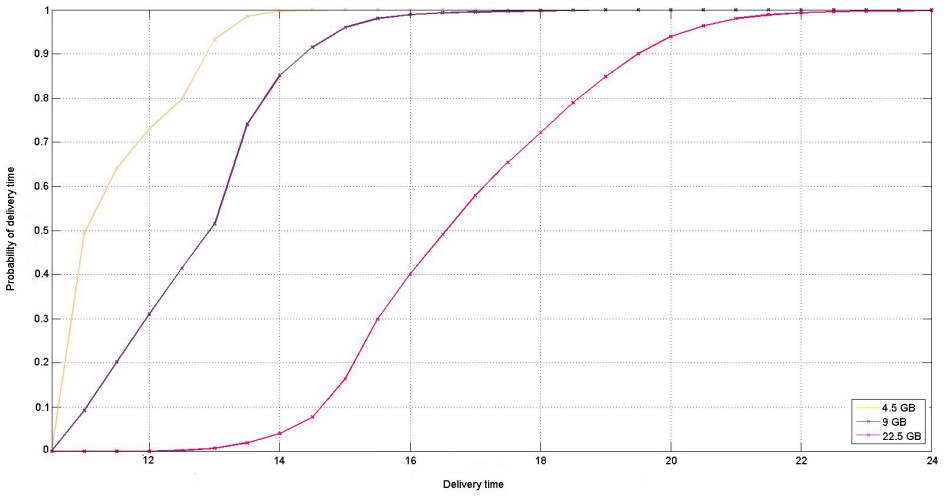

Figure 47: Delivery time probability for 4.5, 9 and 22.5 GB of data, starts at 14:00

6 Conclusion and Future work

Considering the short diary of anticipatory mobile communication in Chapter 2, and comparing it with the experimental results in Chapter 5, it can be seen that the used methodology in this thesis is complimentary for anticipatory mobile networks. The results in Chapter 5 show the ability of user available bits prediction in future based on the analysis of user and network historical data. This ability combined by anticipatory mobile network algorithms, can result in more efficient data delivery. The results of the modeled system in this thesis can be used as input for data segment delivery scheduling algorithms in anticipatory mobile communication, to make scheduling as optimized as possible.

Moreover, this modeled system shows that by predicting the future available data rate and computing the required time to deliver a content with specific amount of data, it is possible to motivate the user through cheaper price or higher data rate, to download his/her required data at time durations in which the bandwidth is less utilized and/or the network load is low.

The assumption in this model is that the network N is serving only one user, and the network dedicated all the available free resources to one user U. In a scenario where more than one user is to be served, network needs to use a smart resource allocation between the users to satisfy the QoS of all the users based on their demanded service, and delivery time efficiency of the data.

One of the issues here is choosing the most optimized sharing method between users. Frequency Division Duplex (FDD) or Time Division Duplex (TDD) are different options. In LTE, it is possible to use a FDD-TDD aggregation system which can provide more bit rate. Based on the used method, system should plan the most efficient division in time and frequency between the users, considering the predicted parameters of the network and users locations. On the other hand, there might be some cases that the ending time of the delivered data is important for the network or the user. In such a case, the network can start the transferring of the data to the user and do it discontinuously to dedicate the resource at some point to more important users, keeping their QoS in the guaranteed level. In other words, the model for such a system, should consider the guaranteed QoS and the demanded type of the data for each user and find formulas to implement these factors in the proposed methodology in this thesis.

Another parameter, which is not introduced in this work's methodology, whose impact can be studied, is the effect of the user's social behavior on the location of the user. A few researches concerning the Location Based Social Networks (LBSN) are performed. By knowing the impact of the social relation of the user on his/her daily temporal Pattern, it is possible to predict location presence probability of the user more accurately.

References

[1] Cisco. Cisco visual networking index: Global mobile data traffic forecast update, 2015-2020. pages 1–39, February 2016. [Online; accessed 30/05/2016], Available: http://www.cisco.com.

[2] Ericsson. Ericsson mobility report on the pulse of the networked society. pages 1–31, November 2015. [Online; accessed 30/05/2016], Available: https://www.ericsson.com.

[3] Veljko Pejovic and Mirco Musolesi. Anticipatory mobile computing: A survey of the state of the art and research challenges. *ACM Computing Surveys (CSUR)*, 47(3):47:1–47:29, April 2015.

[4] OpenSignal. 2G/3G and 4G LTE Cell Coverage Map. Web Application. [Online; accessed 20/04/2016], Available: http://opensignal.com.

[5] Nabajeet Barman and Stefan Valentin. Wireless link quality prediction using street and coverage maps barman. In *Proceedings of 1st KuVS Workshop on Anticipatory Networks*, pages 43–46, September 2014.

[6] Haakon Riiser, Tore Endestad, Paul Vigmostad, Carsten Griwodz, and Pål Halvorsen. Video streaming using a location-based bandwidth-lookup service for bitrate planning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 8(3):24:1–24:19, 2012.

[7] Martin Dräxler, Johannes Blobel, Philipp Dreimann, Stefan Valentin, and Holger Karl. Anticipatory buffer control and quality selection for wireless video streaming. *arXiv preprint arXiv:1309.5491*, 2013.

[8] Martin Dräxler and Holger Karl. Smarterphones: Anticipatory download scheduling for segmented wireless video streaming. In *Proceedings of 1st KuVS Workshop on Anticipatory Networks*, pages 16–19, September 2014.

[9] Keysight Technologies. 4G LTE-Advanced technology overview. [Online; accessed 10/04/2016], Available: http://www.keysight.com.

[10] Tutorialspoint. LTE Quick Guide. [Online; accessed 10/04/2016], Available: http://www.tutorialspoint.com.

[11] Informa Telecoms and Media. Understanding today's smartphone user: Demystifying data usage trends on cellular & Wi-Fi networks. 2012. [Online; accessed 08/05/2016], Available: http://www.informatandm.com.

[12] Qualcomm. The Evolution of Mobile Technologies: 1G 2G 3G 4G LTE. [Online; accessed 01/06/2016], Available: https://www.qualcomm.com.

[13] Ari Banerjee. Big Data & advanced analytics in telecom: A multi-billion-dollar revenue opportunity, Sponsored Report by Huawei. pages 1–24, December 2013. [Online; accessed 18/04/2016], Available: http://www.huawei.com.

[14] International Telecommunication Union ITU. Measuring the information society. pages 1–234, 2012. [Online; accessed 19/04/2016], Available: http://www.itu.int.

[15] Günther Sagl and Bernd Resch. Mobile phones as ubiquitous social and environmental geo-sensors. *Encyclopedia of Mobile Phone Behavior*, pages 1194–1213, 2015.

[16] Venet Osmani, Oscar Mayora, Tinku Rasheed, and Elio Salvadori. Realising anticipatory networks through Human Aware Networking (HAN). In *Proceedings of 1st KuVS Workshop on Anticipatory Networks*, pages 16–19, September 2014.

[17] U. Paul, A. P. Subramanian, M. M. Buddhikot, and S. R. Das. Understanding traffic dynamics in cellular data networks. In *Proceedings of IEEE INFOCOM*, pages 882–890, April 2011.

[18] Hatem Abou-zeid, Hossam S. Hassanein, and Stefan Valentin. Optimal predictive resource allocation: Exploiting mobility patterns and radio maps. In *Proceedings of IEEE Global Communications Conference (GLOBECOM)*, pages 4877–4882, 2013.

[19] Marta C. Gonzalez, Cesar A. Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

[20] Zheng Lu and Gustavo De Veciana. Optimizing stored video delivery for mobile networks: The value of knowing the future. In *Proceedings of IEEE INFOCOM*, pages 2706–2714, 2013.

[21] Kristian Evensen, Andreas Petlund, Haakon Riiser, Paul Vigmostad, Dominik Kaspar, Carsten Griwodz, and Pål Halvorsen. Mobile video streaming using location-based network prediction and transparent handover. In *Proceedings of ACM workshop on Network and operating systems support for digital audio and video*, pages 21–26, 2011.

[22] Eric Malmi, T. Do, and D. Gatica-Perez. Checking in or checked in: Comparing large-scale manual and automatic location disclosure patterns. In *Proceedings of ACM Conference on Mobile and Ubiquitous Multimedia (MUM)*, pages 26:1–26:10, 2012.

[23] Mao Ye, Krzysztof Janowicz, Christoph Mülligann, and Wang-Chien Lee. What you are is when you are: The temporal dimension of feature types in location-based social networks. In *Proceedings of ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 102–111, 2011.

[24] E. Y. C. Wong and Emma Zhou. Assessing factors in mobile marketing context model adopting tam, commitment-trust theory, environment and emotional items

in facilitating purchasing intention. In *Proceedings of International Conference on Service Systems and Service Management (ICSSSM)*, pages 1–6, June 2015.

[25] Huiji Gao, Jiliang Tang, Xia Hu, and Huan Liu. Modeling temporal effects of human mobile behavior on location-based social networks. In *Proceedings of ACM international Conference on information & knowledge management*, pages 1673–1678, 2013.

[26] Huiji Gao, Jiliang Tang, and Huan Liu. Mobile location prediction in spatio-temporal context. In *Nokia mobile data challenge workshop*, volume 41, page 44, 2012.

[27] N. T. An and T. M. Phuong. A Gaussian mixture model for mobile location prediction. In *Proceedings of International Conference on Advanced Communication Technology*, volume 2, pages 914–919, February 2007.

[28] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

[29] Tony Keenan and Rudi Villing. Maximising spectral efficiency in LTE cells. In *ITE Irish Signals and Systems Conference (ISSC)*, pages 1–6, 2012.

[30] P. P. Vaidyanathan. The theory of linear prediction. *Synthesis lectures on signal processing*, 2(1):1–184, 2007.

[31] Wikipedia. Wiener filter. [Online; accessed 14/05/2016], Available: https://en.wikipedia.org.

[32] C. M. Grinstead and J. L. Snell. *Introduction to probability*. American Mathematical Soc., 2012. [Online; accessed 26/02/2016], Available: https://math.dartmouth.edu.

[33] Hossein Pishro-Nik. *Introduction to Probability, Statistics, and Random Processes: Statistics and Random Processes, chapter 3*. Kappa Research, LLC, 2014. [Online; accessed 09/03/2016], Available: https://www.probabilitycourse.com.

[34] Wikipedia. Survival function. [Online; accessed 02/03/2016], Available: https://en.wikipedia.org.